Doctoral Thesis

A reliable past or a reliable pest? Testing canonical stimuli in speech perception research

Johnathan Jones

SN: 17103296

UCL Institute of Education

Primary Supervisor: Dr. Talia Isaacs

Subsidiary Supervisor: Dr. Kazuya Saito

Word count: 56224

Declaration

I, Johnathan Jones confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

A growing body of research is exploring second language (L2) learners' listening perception of vowel contrasts. Conventionally, researchers have estimated how well listeners differentiate between L2 vowels with isolated words (or syllables) in a fixed consonantal frame, such as b-vowel-t (e.g., beat-bit). However, there is a dearth of research that systematically examines how well results generalise beyond isolated frames or the suitability of employing more phonologically and sententially diverse listening prompt types for assessing L2 vowel perception. To address this gap, two studies investigated the effects of using b-vowel-t and more diverse prompt types for assessing intermediate-advanced adult L2 perception of English /i/-/ɪ/ and /ɛ/-/æ/ vowel pairs. Prompt performance was measured for internal consistency, congruence with the Perceptual Assimilation Model for L2 speech learning (Best & Tyler, 2007), and listeners' subjective experiences with each prompt type. Mixed effects modelling investigated the predictive power of b-vowel-t performance on more diverse prompt types. Study 1 explored prompt performance using closed-set, forced choice tasks with first language (L1) Mandarin and Korean listeners. Study 2 investigated the effect of Mandarin and Spanish L1 listeners' target word familiarity and associations with sentence prompts using transcription-response tasks and self-report surveys. Both studies found that diverse prompts had adequate internal consistency and aligned with PAM-L2 predictions. B-vowel-t prompts poorly generalised to diverse prompts and accorded less with PAM-L2 predictions. Survey results showed increased demands from more diverse prompt types based on participants' ratings; however, this did not always correspond to lower performance. Collectively, results indicate utility in employing prompts beyond isolated words in a fixed consonantal frame for laboratory and at-home administrations. These findings contribute to the vowel perception literature by evaluating and extending the scope of prompts which may be used.

Impact Statement

This research examined the use of phonologically and sententially diverse listening prompts to address a scarcity of published literature investigating L2 vowel perception (in advanced learners) beyond isolated words and syllables. Given that vowels contain inherent variability (i.e., their spectral qualities change and are “coloured” by neighbouring consonants) and that speech is rarely heard in isolated contexts, it is reasonable to conclude that assessing vowel perception based on a single, isolated context may underrepresent the construct of vowel perception. It is thus necessary to determine the extent to which perception in a consonantally fixed, isolated context (e.g., b-vowel-t words) generalises to more phonologically diverse and sentential contexts. Further, it is important to uncover the extent to which fixed and diverse prompt types adhere to predictions based on established theory. With these factors considered, diverse prompt types were introduced to (1) uncover the extent to which an isolated consonantal frame generalises to phonologically diverse and sentential contexts, and (2) to uncover the potential suitability of diverse prompt types for empirical inquiry. Results provide insights which may help inform research practice and vowel perception assessment development.

The study found that the isolated b-vowel-t prompt, a commonly used prompt type, has limited generalisability to more diverse contexts and did not always match predictions made based on the Perceptual Assimilation Model for second language learners (Best & Tyler, 2007). Contrastively, the more diverse sentence prompts consistently adhered to predictions. The traditional b-vowel-t prompt type was more internally consistent than the diverse sentences; however diverse sentences were found to hold sufficiently high internal consistency to be used in empirical research.

Results suggest that traditional L2 vowel perception assessments may be unnecessarily underrepresenting the construct of vowel perception. Though sentences increase variability, the variability is arguably relevant to the construct and may be accounted for in analysis through generalised linear mixed modelling (GLMM). GLMM permits the researcher to identify the extent to which individual variables (e.g., association of words with their sentence contexts) impact vowel perception.

Associations offered a promising glimpse into the interaction between top-down and bottom-up processes, a confluence identified as important to the construct of listening perception (Field, 2004). This licenses further investigation to identify associations between these processes and assimilation types (e.g., PAM-L2). There may be within- and between-assimilation-type gradations that emerge from incorporating prompts which explore these processes.

Lastly, this work has potential implication for high variability phonetic training, a natural extension for the present research. Though results from the present study would suggest that training on individual, isolated contexts should not be expected to generalise to diverse sentential prompts, it is possible that training using such prompts might. Thomson (2012) noted that unbounded variability would be detrimental to perception, but in the context of connected speech prompts, how much variability and for which target group (e.g., a given language, age, or proficiency) are relevant investigations. It may be that for some groups (e.g., more advanced learners) or contexts, the added variability yields better training results than isolated speech prompts.

Dedication

This thesis is dedicated to my amazing wife, Hwajung Monica Nam. Thank you for the love, support, and patience you have selflessly given throughout these studies and beyond. Hopefully this thesis will wedge open doors for us which have until now remained closed.

Acknowledgements

많은 사랑과 응원을 보내주신 아버님, 어머님께 진심으로 감사드립니다. 두 분께서 저와 화정이를 위해 해주신 모든 것들, 결코 잊지 않겠습니다.

To my primary research supervisor, Talia Isaacs, thank you for your guidance, care, and unwaveringly accessible support, rain or shine. Your keen acumen and timely assistance have been cherished. To Valerie Hazan, my initial subsidiary supervisor who graciously committed to continuing supervision even after retirement, thank you for your time and sagely advice in shaping this study. Best wishes in retirement! To my subsidiary supervisor Kazuya Saito, who took over from Valerie as subsidiary supervisor at a very late stage. Thank you for taking on my study despite a whirlwind schedule. Your help advanced my thesis as well as my academic perspective. Thank you to John Gray for his help as interim supervisor in a maternity cover. Thank you to my examiners, Murray Munro and Chloe Marshall for your perspective and the opportunity to defend my work.

Thank you to those who helped along the way. To Michael Rochemont. I will always remember your dedication, integrity, and compassion with admiration. Rest in peace. To Douglas Pulleyblank, Rose-Marie Déchaine, and the excellent UBC Linguistics Department. To Scott Mackie for advice and assistance in all things phonetics. To Carsten Roever, Tim McNamara and Paul Gruba, thank you for the knowledge and opportunities you provided. Thank you to Nathanael Lambert for nurturing my academic aspirations during my Masters, introducing me to my first doctoral reading group, and welcoming me to the doctoral study room. You considered me one of the team before I was even ready to aspire to be. Arpan Tahim, Jumana Al-Waeli, Ketan Dandare, Lorena Sanchez-Tyson, and the group at UCL IOE, it has been a privilege. You were an endless source of smiles and comfort. Thank you to Gwen Brekelmans for helping me break through numerous roadblocks. Thank you to Zhi Li and Judith Fairbairn for your qualitative coding and assistance beyond.

And finally, thank you to my wife, family and friends. It is only through your love and support that I was able to get to this point. I love you dearly.

Table of contents

List of tables

List of figures

Glossary

Terms and abbreviations used in this thesis:

2 alternative forced choice, 2AFC (also 3 or 4 alternative forced choice)	The number of options a listener must choose from to identify the correct response in a listening task.
Allophone	Instances of a consonant or vowel which are physically different, but perceived as the same. For example, the /p/ in *port* and *sport* are perceived as the same to an English L1 speaker, but are physically different (word initial /p/ is aspirated while /p/ after /s/ is unaspirated. These two p's are non-contrastive in English, but contrastive in other languages.
Assimilation	How an L2 listener perceives or organises an L2 vowel or consonant. According to the Speech Learning Model and Perceptual Assimilation Model, L2 segments are assimilated according to similarity and dissimilarity with L1 segments.
Block	A grouping of prompts administered by prompt type. For instance, the b-vowel-t (bVt) oddity block is a section of strictly bVt words (rather than diverse words or sentences) in an oddity discrimination paradigm.
bVt (b-vowel-t); bVd (b-vowel-d)	Isolated words with a word-initial /b/ followed by a target vowel and /t/. The bVt frame was used in the two primary experiments, while bVd was used in the pilot. To elicit participant feedback, bVt was labelled b-vowel-t to facilitate participant understanding. Consequently, direct quotes from participants read "b-vowel-t" rather than "bVt".
Closed set response type	Used in Experiment 1 and 2, response options in a task are given to participants. Answers in a task are limited to these options. For identification tasks, this can be the choice between two minimal pairs; in an oddity task this is a choice between one of four options. This is contrasted with an open response type.
Cognate	One of two vowels or words which constitute a minimal pair. In the minimal pair, bet-bat, bet and bat are cognates.
Construct	The concept of what is being measured or assessed. The construct definition provides the basis for developing an assessment, including its items (e.g., listening prompts) and tasks. Vowel perception and what it entails is the construct examined in the present research.
Construct irrelevant variance	Variance in participant performance (i.e., scores) that is not due to the construct being assessed. This is analogous to confounding variables where an uncontrolled variable or variables are responsible for results.
Construct underrepresentation	Not enough of the construct is being assessed. This can lead to an inaccurate or incomplete understanding of participant ability, such as an L2 listener's ability to differentiate between L2 vowels.
Discrimination task	A task which requires listeners to respond to differences in a word or sentence (e.g., to indicate that the third word is the odd word in the sequence, "bet-bet-bat"). Identifying what the precise vowel, word or sentence is is unnecessary in discrimination tasks. The present study employs 3- and 4-interval oddity discrimination tasks.

Diversity and complexity	Variations in the phonological environment (e.g., neighbouring consonants), syllable type (e.g., mono- or disyllabic), or syntax (e.g., place in a sentence) that a target vowel resides. The term is used relatively in the present research to compare various prompt types (e.g., Diverse Sentences) with isolated bVt or bVd words. Diversity may be used interchangeably with complexity or speech signal complexity. However, diversity as a label was selected over complexity to avoid ambiguity. Labelling diverse sentence prompts as complex sentence prompts (opposed to carrier sentence prompts, which have a fixed syntax and word-final placement of target words) could predictably and unnecessarily be mistaken for sentences containing embedded clauses.
Fixed frame, fixed consonantal frame	A listening prompt used for vowel perception research which consists of a single consonantal environment, such as bVt, where the vowel may change across trials, but the consonantal context will not.
high (front) vowel pair	The vowels /i/ and /ɪ/. High and front indicate the position of the tongue in the mouth as it articulates the vowel. As front is given for target vowels in this research, it is dropped for parsimonious reading.
Identification task	A task which requires listeners to identify the word or sentence they heard. In the current study, this is done by either mouse-clicking an on-screen button or by transcribing what was heard.
Implicational hierarchy	A unidirectional hierarchy where the presence of given component implicates the presence of another, but not vice versa. For example, having £10 implicates having £9, and having £9 implicates having £8, but having £8 does not implicate having £9.
Interval	The number of stimuli presented in a listening task. A three-interval oddity task indicates there are three words presented in sequence.
Mid-low (front) vowel pair	The mid (/ɛ/) and low (/æ/) front vowels. Mid and low indicate the position of the tongue in the mouth as it articulates the vowel. As front is given for target vowels in this research, it is dropped for parsimonious reading.
Oddity	A discrimination task which requires a participant to identify the odd stimulus in a sequence of stimuli (e.g., the word sequence “beat-bit-beat”, where bit is the odd stimulus).
Open response type	Used in Experiment 2, options are not explicitly provided for participants. Participants listen and write what they hear. This may be an isolated word or connected speech. Note that this is a modified “open” response where options are expected to be limited by the number of desirable options available. For instance, where an isolated prompt is the word, “bit”, a listener may be expected to tend to hear “bit” or “beat”, but may also hear words such as “bet” or “but”.
Perceptual constancy	The ability to perceive a stimulus (e.g., a vowel) as that stimulus and not another in a variety of contexts.
Predictive efficacy	The extent to which participant performance on a prompt type (i.e., bVt) is able to predict performance on another prompt type (e.g. Diverse Sentences in Experiment 1 and 2). Efficacy in this study is measured through generalised linear mixed modelling.

Stimulus	A listening prompt used to assess vowel perception. Listening prompts may be an isolated CVC word, diverse words, or sentences.
The generic (street name)	the type of street (e.g., road, lane, street, etc.). The generic is used in conjunction with the “specific” to make a complete street name.
The specific (street name)	the identifying name of a street (e.g., Allen, Redcliffe, Simmons). The specific is used in conjunction with the “generic” to make a complete street name.
Target vowel	The vowel of interest. The vowels of interest in this research are /i/ and /ɪ/, /ɛ/ and /æ/.
Target word	The word in which a target vowel is embedded.
Trial	a discrete stimulus or sequence of stimuli which requires the listener to react in an observable, measurable fashion. The “bit-beat-bit-bit” sequence is an example of a trial which requires participants to indicate where the odd word resides in the four-word array.

Introduction

When vital information is communicated orally, perceiving speech correctly can be critical for making a felicitous decision. Meeting an associate at Radcliffe Square, for instance, may easily be mistaken for meeting at Redcliffe Square, or someone not wanting to live may be understood as not wanting to leave. In such situations, social, environmental, and linguistic context may not be enough to disambiguate the meaning before a decision is made. Speakers of English as a second language (L2) are particularly susceptible to conflating vowels and consonants (segments) of the second language (Thomson, 2017), regularly doing so in instances where no contrast between the L2 segments exists in their first language (L1). Conflating segments which have a high functional load (e.g., they contrast many words) can have a snowball effect when more than one instance is encountered in an utterance, compounding the threat to intelligibility and comprehensibility (Munro and Derwing, 2006). Accurately interpreting a speaker's utterance is important regardless of stakes; however, in professional environments with linguistically diverse populations, it becomes particularly salient, and consequently there have been recent calls for workplace-related language and pronunciation training, whether in the general workforce (Derwing et al., 2014) or in professional spheres such as the field of medicine (Khan, 2016; Yager, 2016). It is within this context—advanced-level, ambiguous speech for English for Specific Purposes—that the present research study is conceived.

Being able to effectively assess and improve L2 learners' ability to distinguish between contrastive segments with high functional load, such as vowels, is desirable to help English language learners interpret otherwise clear language, but what entails "effective" is arguably unclear. Current assessments help indicate whether an L2 speaker can differentiate between two segments, yet do so in strictly attenuated phonological conditions, such as isolated syllables or words in a fixed consonantal frame (Jones, 2015). This leads to binary results with uncertain generalisability. Results can answer, "Can a listener perceive the difference between two sounds in their L2 at a desirable threshold?" but not how well or under what conditions. By adding phonological and sentential diversity to listening prompts, it may be possible for more nuanced information to be gathered about the listener. For instance, an assessment that uses a bVt consonantal frame may identify whether a listener can perceive the distinction between /ɛ/ and /æ/ in the words “bet” and “bat”, yet it would not be able to indicate whether the listener would be able to do so in connected speech, particularly where the word is not immediately associated with the context.

For reasons to be discussed in the literature review, speech perception assessments have historically employed isolated, fixed consonantal prompts (e.g., bVt, hVd, hVbə) to investigate vowel perception. Investigating the use of bVt, a commonly used prompt type, compared to more phonologically and sententially diverse prompt types represents a fundamental, novel contribution to the study of L2 vowel perception. In contrast to historical approaches to L2 vowel perception assessments where homogeneity of prompts is sought, this doctoral research views listening prompt heterogeneity as a potential source of new information, providing a deeper, more accurate and robust picture of a participant’s ability to discriminate between targeted speech sounds.

The study is framed such that it motivates the inquiry through a review of the literature, describes a pilot which helped hone the prompts and methods used in the primary research, and reports two studies which investigates the use of phonologically and sententially diverse listening prompts for assessing L2 vowel perception. Results are gathered and discussed in a general discussion.

Literature review

After a period of decline in the late twentieth century, the twenty-first century has seen pronunciation-related research become a resurgent, interdisciplinary domain of interest (Jones & Isaacs, 2022). Much work has been done to establish pronunciation as an important scholarly pursuit, with contemporary work centring around intelligibility, comprehensibility, accent, and interactions between them. The study of pronunciation may be split into affiliate interests—perception and production—where the bulk of inquiry is focused on production (Monteiro & Kim, 2020). With less output relative to production and even fewer studies related to the processes involved (Field, 1999), there exists an ongoing need to reinforce the literature in listening perception. The current study examines a granular component of listening perception—vowel perception—investigating its means of assessment.

This section highlights relevant elements which motivate the exploration of vowel perception assessment designs. The structure of the literature review includes establishing the basic construct of vowel perception, adopting a theoretical framework to explore perception in an L2 context, and delineating how vowel perception is typically assessed, thereby uncovering the theoretical gap which the present research helps fill: providing empirical support for the implementation or rejection of different types of listening prompts in vowel perception assessments.

L2 vowel perception assessment design

L2 vowel perception assessment is part of a longer history of L1 speech perception dating back a century (see A brief history of vowel perception assessment). As will be discussed later in the literature review, many of the same tenets and methods used in constructing the original L1 assessments remain in use by designers of L2 perception assessments. There have been advancements in assessment design theory, however, which have helped inform perception assessment designs, but have not all made their way into L2 vowel perception contexts. Such advancement includes explicitly defining the construct (i.e., what precisely is being assessed) and the corresponding means of assessing it (Cronbach & Meehl, 1955).

For the present study, the most informative, single document for speech perception design was Bilger (1984). Bilger, a heavily cited pioneer in speech perception research for assessing hearing impairment, identified a divide between speech perception assessment and contemporary psychological assessment principles. In summarising design methods, Bilger explained “primary importance” should be placed upon defining the construct (p. 2). Bilger noted that designers of commonly used speech perception assessments did not historically have the benefit of more recent advances in assessment theory (i.e., Cronbach & Meehl, 1955), and that their constructs were inferred as opposed to explicitly defined. Inferences about the construct of these assessments were made by looking at “observables” such as the prompts and tasks. Because speech perception assessments (or the prompts they employ) had largely been based on designs for evaluating radio and telephone systems rather than speech perception, the assessments may not necessarily reflect their intended use. In other words, listening prompts should be developed in a manner which reflects the construct they are used to assess rather than simply complying with historical practice. Though progressive at the time, specifying the construct and the intended use of an assessment instrument remain important considerations in modern designs (Bachman & Palmer, 2010; Chapelle, 2020; Kane, 2016).

Discussing the construct and how it tended to be represented, Bilger identified common characteristics of speech perception assessments, noting the use of isolated, monosyllabic words spoken by “typical talkers” in “atypically precise” language (p. 3). Assessments which employed such prompts, Bilger reasoned, likely underestimated a listener’s difficulty in understanding speech. The instrument did not adequately assess vowel perception or properly inform decisions about a listener’s ability. While this claim was in reference to hearing impairment, the same reasoning holds for L2 speech perception. By employing restricted, isolated prompts, a researcher may risk overestimating a listener’s skill at perceiving speech in the real world, jeopardising the potential to identify a salient need.

To avoid such outcomes, Bilger claimed “any designer of speech recognition tests should devote some thought to the construction of items that are realistic samples of speech” (p. 3), with sentences constituting more realistic speech than isolated words. Though not stated explicitly, this refers to authenticity, which will be discussed in a later section.

Assessment theory (particularly language assessment) has evolved since Bilger to include greater emphasis on validity and authenticity, as well as incorporating participant experience and interactivity (Bachman & Palmer, 1996, 2010; Canale, 1987). Guided by these concepts, the literature review will now explore the basic construct of vowel perception.

Building the construct of vowel perception

To identify construct relevant information for prompts to explore, it is important to firmly establish what vowel perception entails. The subsequent sections highlight characteristic features of vowels and discuss processes involved in how they are perceived by L2 listeners.

Vowel characteristics. In vowel production, the harmonically-rich spectrum of the glottal source is filtered as a result of resonances in the supralaryngeal vocal tract. This interplay results in spectral peaks called formants that appear in the speech output. Vowels are a combination of several resonant frequencies called formants. The first (F1) and second (F2) formants are primarily responsible for the vowel's identifiable features (i.e., its quality). F1 is inversely associated with height, where a low F1 corresponds with a higher vowel. There are articulatory, acoustic, and auditory definitions of height, but for simplicity, vowel height pertains to height of the tongue in relation to the mouth[1]. In English, the vowel /i/ is denoted as the highest vowel and has the lowest F1, whereas /æ/ is the lowest vowel and has the highest F1 (Hillenbrand et al., 1995). F2 is associated with backness, where back vowels are articulated with the tongue body toward to back of the mouth, while front vowels (a lack of backness) are produced with the tongue body toward the front of the mouth. Higher frequency F2 is more associated with backness, and lower frequency F2 is related to less backness. Vowels produced with the tongue at the back of the mouth will have a lower frequency F2 compared to a vowel produced with the tongue towards the front of the mouth. Standard Southern British English (SSBE) contains four front vowels, /i/, /ɪ/, /ɛ/, /æ/ (Evans & Alshangiti, 2018). Front vowels are the centre of much inquiry due to how they are perceived by individuals from different language backgrounds and are the subject of inquiry in the current research. A final frequency band of interest is the *fundamental frequency* (F0), which is responsible for what is heard as pitch (Ladefoged & Johnson, 2014). Distinctions in F0 are audible, but are "category independent" (Hojen & Flege, 2006, p. 3073). In isolated contexts and when spoken by the same talker, F0 may help a listener distinguish two juxtaposed vowels.

Spectral properties, particularly from F1 and F2, are primary indicators of vowel identity, but vowels are also accompanied by durational cues which, in English, act as secondary identifiers

[1] Pulleyblank (2011) explains that for illustrative or pedagogical purposes, F1 can be thought of as corresponding inversely with the height of the tongue, but that in strict mapping of articulation to acoustics, the relationship occurs "only sometimes" (p. 492). For instance, in advanced tongue root languages, the tongue position for /e/ can be higher than /i/.

(House & Fairbanks, 1953). Though vowels are associated with intrinsic lengths, duration is non-contrastive in English and is subject to overlap in different contexts (Peterson & Lehiste, 1960). Hence, if a listener who uses length as a primary cue for identification, as L2 learners tend to do (Kim et al., 2018), the listener may be prone to misinterpreting an utterance where the target vowel plays an important role in understanding an interlocutor's message.

The explanation thus far may give the impression that vowels are fixed entities with fixed spectral qualities; however, vowels are characteristically variable, stemming both from endogenous (vowel inherent spectral change and allophonic variation) and exogenous (neighbouring consonants) factors. Vowel formant frequencies are inherently variable, and it is within that variability that vowels are most intelligible (Hillenbrand, 2013). Intelligibility as a function of a vowel's variability has been identified as a *dynamic view*[2] (Morrison, 2013), a relatively new concept that may be contrasted with the *static view*. Until recently, it was thought that a vowel's steady state, where they are least spectrally variable, displayed the vowel's true quality (Williams et al., 2018). Thus, the steady state, often toward the centre of the vowel is where measurements were made. Rosenblum (2008) explains, "Research shows that silent-center syllables are recognized as easily as intact syllables and that the extracted portion of the syllable, which should contain the most "canonical" portions of the vowel, are relatively less informative". The vowel's quality, therefore, is in its characteristic variability, not its fixed state.

Beyond vowel inherent spectral change, there are different versions of each vowel brought about by changes in neighbouring consonants (Levy & Strange, 2008). Functionally, a vowel is not a single sound (phone), but a group of sounds and associated features which are *perceived* as a single sound (phoneme). To illustrate, the sentence, "Bees keep busy preening", contains four instances of the vowel, /i/, each being unique temporally or qualitatively. The /i/ in "keep" will be temporally shorter than the /i/ in "bees"[3] and the /i/ in the second syllable of "busy" is shorter still because it is

[2] The dynamic view of vowel intelligibility is unrelated to dynamic systems theory.
[3] Vowels are shorter before voiceless obstruents (e.g., /p, t, k, f, s/) compared to voiced obstruents (e.g., /b, d, g, v, z/).

in an unstressed syllable. Consonants neighbouring the vowel alter the vowel's spectral qualities. The /i/ in "preening", for example, is nasalised as the speaker's velum lowers in anticipation of the /n/ and "colours" the /i/ with nasality. Such variations are non-trivial as certain language groups use length (e.g., Thai) or nasalisation (e.g., French) contrastively to distinguish between different words, while in English, each of these instances of /i/ are perceived as the same vowel (Krämer, 2019). In English, these differences constitute non-contrastive, within-category variation. Hence, within-category variation is any non-relevant difference within a person's internal construct of what a given vowel is (e.g., different allophones, pitch or amplitude), opposed to between-category variation which distinguishes contrastive sounds (e.g., F1 and F2 in vowels).

In an L2 context, there is evidence that differences in the vowel stemming from neighbouring consonants may affect a listener's perception (Strange et al., 2001). For instance, Levy and Strange (2008) found that inexperienced learners of Parisian French had more trouble discriminating between /i-y/ in a bilabial context (rabVp) and more trouble with /u-y/ in an alveolar context (rabVt). The experienced group had difficulty with both contexts. Overall, the inexperienced learners discriminated better with rabVp than rabVt, whereas the experienced learners displayed no significant effect for context. Differences in neighbouring consonants is addressed in this study through analysis, where differences in listening prompts are part of the item's error (see Analysis and statistical approach); however, endeavouring to uncover the relative effects of each environment were beyond the scope of the present research.

With the variable nature of vowels described, the cognitive processes involved in their perception will now be discussed.

Bottom-up and top-down processing. Listening information processing is often bifurcated into top-down and bottom-up. The relationship between these processes is that between the phonetic features of an utterance and its overall meaning, where meaning is derived from knowledge of the world and speaker, from schemas which cue expectations, and from associations stemming from words in a sentence (Field, 1999). L2 listeners employ top-down and bottom-up processes jointly and automatically, regardless of proficiency level (Field, 2004). Field (1999) notes that in the context of L2 listening, it is commonly held that less proficient learners are more inclined to use lower level, bottom-up processes opposed to focusing on meaning; higher levels learners typically use a more top-down approach which focuses on meaning. Augmenting this claim, Field cites Stanovich (1980), explaining, "we use contextual information to make up for unreliability in the signal (bad handwriting, for example, or ambient noise). The more flawed the bottom-up information, the more we draw upon cues from top-down sources" (p. 339). In an L2 context, listeners may have difficulty differentiating between certain contrastive vowels (i.e., two distinct vowels are perceived as homophonous by the L2 listener). In such cases, the "flawed" information is the homophony of the difficult vowels. Where the vowels are homophonous, words which are contrasted by the vowels (e.g., bed, bad) are homophonous, and therefore top-down processes are engaged to resolve the homophony (Halberstadt et al., 1995).

In sentential contexts, the process of resolving homophony can be conceived as part of a *semantic network*, such as that posited by Collins and Loftus (1975) and applied contemporarily in cognitive sciences (e.g., Darcy et al., 2012; De Dayne et al., 2017). A semantic network is a theoretical model of speech information processing, encompassing all words (ideas), interrelations between ideas, and the relative distances between those ideas. Each word in the sentence activates associations, spreading to other words and ideas ("spreading-activation"). Spreading-activation applies to lexical, syntactic, and phonological categories; associations can stem from each of these. For cognitive economy, more frequent words are accessed more readily than less frequent words (i.e., they are "nearer"), and more closely related ideas are accessed faster than less closely related ideas. Hence, where a listener cannot discriminate well between two L2 vowels, that listener would be expected to perceive the word most strongly associated with the sentential context. Where word frequency and syntax are controlled and the listener has no association for one word over another in a given sentence, the listener would be expected to perceive the phone in a manner most salient to him or her, that which most closely aligns with the listener's L1 phonology. A hypothetical Mandarin L1 listener, for example, has the L1 category[4] /ɛ/, but not /æ/ (see Selecting a model of L2 perception), and has difficulty differentiating between these English vowels. Should the listener hear the sentence, "I said, 'bat'", and not have a stronger association of bat than bet in the sentence, the listener would be predicted to perceive /ɛ/ (bet) rather than /æ/ (bat). Such semantic and lexical equivalence is not often present, however, and thus an interaction between bottom-up and top-down can be anticipated.

Next, how vowels are perceived in an L2 context will be discussed, starting with selecting an appropriate model for the purpose of this research.

Selecting a model of L2 perception

Several models exist which help explain how L2 speakers perceive and acquire (assimilate) L2 segments[5], but the two that have been the most influential (Bohn, 2017; Tyler, 2019) are the Perceptual Assimilation Model (PAM; Best, 1995) and the Speech Learning Model (SLM; Flege, 1995a). Both have revised versions which will be discussed later in this section.

L2 perception in PAM and SLM is built on a common premise that L1 phonology influences L2 phonology, and L2 phones which are similar to an L1 category can be more difficult to master than phones which are dissimilar. PAM was designed to explain L2 listeners' speech *perception* for naïve learners, making "explicit predictions about assimilation and discrimination differences for

[4] Here, category refers to the mental representation of a vowel that is heard.

[5] The interested reader may consider the Automatic Selective Perception model (Strange, 2011), Natural Referent Vowel framework (Polka & Bohn, 2011), Native Language Magnet theory (Kuhl, 1992) and the Native Language Magnet theory expanded (Kuhl et al., 2008). Also, Bohn (2017) provides an excellent summary of L2 segmental mapping.

diverse types of non-native contrasts" (Best et al., 2001, p. 777). SLM was intended to account for L2 speech *production* learning over a person's life span (Flege & Bohn, 2021). Rather than contrasts, SLM targets individual sounds. The model posits a direct correspondence between perception and production, where accurate perception is a fundamental precursor for accurate production. Because production stems from perception in SLM, it can also effectively explain perception (Bohn, 2017, p. 223). This is another dissimilarity with PAM, as in PAM "accented speech is not necessarily an impediment to the acquisition of new L2 categories" (Tyler, 2019, p. 616).

A crucial distinction between the models is the type of categorisation they target. SLM accounts for phonetic-level categorisation while PAM encompasses both the phonetic and phonemic categorisation. A strictly phonetic interpretation entails that each L2 speech sound (phone) is assimilated individually. It is context sensitive, meaning perception in one environment (e.g., bVt) may not translate to another (Thomson, 2012). While addressing individual phones and their effects for specified groups of listeners or individuals is a valuable inquiry—an acknowledged avenue for subsequent exploration—it is beyond the scope of the present paper. As indicated in the section, Vowel characteristics, this paper has posited a phonemic perspective and is primarily concerned with between category variation (information used to distinguish between vowels) rather than within category variation (non-contrastive variation within a vowel). A phonemic interpretation encompasses not only a single phone, but the group of phones (allophones) which constitute a person's mental representation of the segment. Where investigating perception beyond a single, isolated consonantal frame, there will necessarily be various instances of a target vowel. Described in Vowel characteristics, these instances will vary physically while still representing the same category. Consequently, a phonemic model is more appropriate for a study that investigates perception in varied contexts, suggesting PAM is better suited for the present research than SLM. However, PAM targets naïve rather than experienced learners, making SLM more appropriate. A resolution is found in PAM-L2 (Best & Tyler, 2007), the revised version of PAM. Building from the assumptions of the original model and with the backing of empirical research, PAM-L2 generalises its predictions to L2 learners' experience over time in an L2 environment (Tyler, 2019). As PAM-L2 targets phonemic-level perception of more advanced learners, PAM-L2 was the natural choice for a theoretical framework for the present study.

For reference, SLM was recently revised, becoming the SLM-r (Flege & Bohn, 2021). Yet it retained its phonetic emphasis, making both the SLM and SLM-r inappropriate for a study which incorporates varied contexts.

Having determined the appropriate model of perception, it is now pertinent to delineate its predictions. In PAM-L2, L2 phones are assimilated into a speaker's native phonetic and phonemic inventories based on the extent of similarity to L1 categories. PAM-L2 predictions for category formation are based on those described in PAM (Tyler, 2021). If two non-native phones are assimilated into two distinct phonological categories (TC), the model predicts discrimination will be "tantamount to discrimination of a native contrast" as "it is native-language phonological attunement that is responsible for the accurate discrimination" (Tyler et al., 2014, p. 5) . If two phones are assimilated to the same phonological category, discrimination is expected to be poorer; how much poorer depends on how well each vowel exemplifies the category to which it is assimilated. Two phones which are considered equally good or poor representations of a native category are predicted to be harder to discriminate between than when one phone is considered a good representation while the other is poor. The former is single-category discrimination (SC) while the latter is category-goodness discrimination (CG). If the phones have no native category which they may assimilate to, they remain uncategorized. A single uncategorized phone is expected to contrast well with categorized phonemes (UC), while two uncategorized phones (UU) may lead to poor to excellent discrimination, depending on the phonetic similarity the phones share with each other and a given native phonetic category. UU does not describe the current dataset and will not be discussed beyond this section. A discriminability hierarchy, from easiest to most difficult assimilation pattern, has been posited as TC/UC > CG > SC (Tyler et al. 2014).

Assessing vowel perception

A brief history of vowel perception assessment. Listening prompt designs in vowel perception assessments have changed relatively little over the last century, with isolated syllables and words persisting as the incumbent means of investigation, and connected speech predominantly unutilised. The reason specified for employing isolated words has evolved over the last century, but given the literature, does not appear to have been formally and cohesively documented. Uncovering this history pours the foundation for the present study.

Modern speech perception testing methodology traces back over a century to where Bell Telephone Laboratories (Bell Labs) conducted research to explore and improve intelligibility of their speech transmission systems (e.g., Campbell, 1910; Crandall, 1917; Fletcher, 1929). The texts, *Speech and Hearing* (Fletcher, 1929) and *Articulation Testing Methods* (Fletcher & Steinberg, 1929), consolidated methods for testing intelligibility[6] and established standards still used in modern speech perception research. Relevant to the current study, these texts summarised features for syllable, word, and sentence list creation and outlined separate criteria for intelligibility depending on the list employed.

Fletcher's pioneering Speech and Hearing suggested two measures for operationalising intelligibility, beginning with the perception of vowels and consonants in simple syllables (e.g., "wa", "yip") and words (e.g., "bought", "bit"). Listeners heard these speech sounds and manually transcribed what they heard on a response sheet. Intelligibility was determined by the percentage of consonants or vowels correctly perceived, with correct perception inferred by correct transcription. A second measure of intelligibility came through using sentences. Criteria for sentence creation dictated that sentences should test the observers' perceptual acuity while minimizing the demands on intelligence. Fletcher provided several motives for using simple constructions when recording speech for testing purposes. Simple syllables were recommended when it was difficult to train

[6] The research was predominantly L1 at the time, but would apply to either L1 or L2 communication.

participants to pronounce or transcribe speech sounds without error, when efficiency was paramount, or when potential memory effects were a concern. A final constraint noted by Fletcher was that psychological factors may affect responses to sentences. Though Fletcher does not explain what he meant by "psychological factors", the term psychological factors (also written as "psychological aspects") was used in other studies of the era when referencing cognitive influences in judging task fulfilment (e.g., Richardson, 1940, p. 842). Thus the confounding psychological factors were potentially introduced by the rater or examiner rather than the listener.

Future Bell Labs research focused on testing segments rather than sentences. Formative works such as "Toward the specification of speech" (Potter & Steinberg, 1950) and "Control methods used in a study of vowels" (Peterson & Barney, 1952), applied Fletcher's techniques for conducting vowel perception experiments, including utilizing monosyllabic words and controlling for possible listing effects. Each method was introduced to minimize variability and ensure what was intended to be tested was in fact being tested.

The goal of these post-war studies was to specify identifiable characteristics of individual segments of speech, both acoustic features and theoretical assumptions. A fixed consonantal environment, /hVd/, was often employed as listening stimuli (e.g., Bogert, 1953; Peterson & Barney, 1952, Potter & Steinberg, 1950); however, the justification was distinct from Fletcher's. Rather than employing the fixed frame for practicality, it was adopted to obtain a "practically steady state" (Peterson & Barney, 1952, p. 177), reasoned to reflect the vowel as it was intended by the speaker or a vowel's most characteristic representation (Lehiste & Peterson, 1961)[7]. Contemporarily, this has been considered a *static view* of vowel perception (Hillenbrand, 2013; Morrison, 2013). To their credit, researchers of the period—at least those indicated above—understood that focusing on a

[7] Fixed consonantal frames such as hVd, sVk, bVd, and pVl had previously been used for vowel comparisons as they form separate words for illustrative purposes (Firth, 1935). Consequently, it would be reasonable to presume that word formation was one of the purposes for selecting hVd as a frame. However, neither Peterson and Barney (1952) nor other vowel perception studies at the time explicitly stated word formation as a selection criteria for implementing the hVd frame.

vowel's steady state was reductivist[8], postulating that connected speech offered complexities that might not be adequately explored through examining a cross section of speech sounds. Potter and Steinberg (1950), for instance, observed that, "if we can specify the vowels in terms of acoustic measurements, we will be in a position to extend the particular measurements that appear significant for specification to the more complex speech situations" (p. 807), indicating that connected speech (or at least being able to generalise toward it) was the ultimate target.

More recently, speech perception researchers have noted that vowels are not only inherently variable due to spectral changes, but that the fluctuations in each vowel lead to greater perceptual accuracy than the isolated steady state of the vowel (see Variability). It is precisely this susceptibility to variation that the present study questions the strict use of fixed consonantal frames for assessing vowel perception (see Validity).

Today, fixed consonantal environments remain canonically employed in vowel perception testing, whether /hVd/ (Loakes et al., 2017), /bVd/ (Barreda, 2017), or other variations of a fixed frame. Whereas in the 1950s a steady vowel state was thought to be most conducive to perceiving a target vowel, present usage of the fixed frame environment, specifically those implementing /h/ or /b/ as neighbouring onset (i.e., word- or syllable-initial) consonants, argues that such consonants minimize the effect of coarticulation on the vowels (Bundgaard-Nielsen et al., 2011; Strange et al., 2007). This follows studies such as Strange et al. (2001) and Levy and Strange (2008), who found that…

This recent justification represents a subtle, yet distinct shift from intending the vowel to be more comprehensible to the vowel being less susceptible to unintended variance in the speech signal. As happened 50 years prior, the shift provides an updated justification while maintaining the historical usage of fixed consonantal frames over diverse words or sentences. The changes in the literature, however, are predicated upon the assumption that limiting variability for exploring vowel

[8] It was theoretical for scholars of speech perception, but not necessarily for those in other fields, such as those interested in pedagogy or general acoustics (e.g., Richardson, 1940). Such researchers believed spectral changes were not important in the perception of speech, at least at the time of reference.

perception is more reflective of the construct than incorporating more variability. It is not predicated on evidence which shows added variability is deleterious for empirical inquiry.

Despite the shifts in justification which enable a seamless continuity of practice for the fixed isolated frame, there is evidence that the diverse, contemporary field of speech perception research is ready to bend, if not break, tradition. Thomson and Derwing (2016) write, "that within-category variation is natural suggests that teaching the pronunciation of L2 sounds should incorporate and emphasize variation rather than focusing on elusive prototypes, citation forms, and the pronunciation of sounds in isolation" (p. 89). This view has manifested itself in high variability phonetic training, the topic of the next section.

High variability phonetic training (and assessment). One of the most influential advancements in speech perception testing and training is high variability training (HVPT), a computer-mediated[9] technique where learners are presented with natural speech exemplars produced by various speakers. While HVPT is a training paradigm, an integral component to most reported HVPT studies is generalisability testing (Thomson, 2018). Without effective testing, claims of HVPT utility are limited. I know of no research which examines prompt types used for assessments, making HVPT a natural venue for the present research to help inform practice.

The first HVPT studies were published as a series by Logan et al. (1991)[10] in response to poor results from "low variability" training studies, specifically Strange and Dittmann (1984). Strange and Dittmann found that training Japanese participants to perceive the English /r/ and /l/ contrast using synthetic speech was moderately successful, but that results did not generalise to natural speech samples (words produced by human talkers). Logan et al. hypothesized that the synthesized speech samples Strange and Dittmann used for training participants were fundamentally impoverished, that

[9] A precursor to the computer mediated training can be seen in identification experiments conducted by Edman and Soskin (1977), where participants dVd or V syllables spoken by 5 talkers were able to learn and generalise more effectively than a control. Unlike HVPT, however, "training" simply included repeated exposure rather than formal training procedures or feedback. Researchers surmised improvements were due to vocal tract normalisation, task learning, and perceptual learning.

[10] These had yet to be labelled HVPT, a term later coined by (Iverson et al., 2003).

synthesized speech provided insufficient acoustic information to generalise to natural speech. Consequently, Logan et al. revisited Strange and Dittmann's study to see whether adjustments to the methodology would lead to better results.

Challenging Strange and Dittmann's findings, Logan et al implemented a number of changes in their study: human speech was used in training, multiple talkers were used to record the speech, the task for training was the same task for testing, and the testing and training words exclusively used words which started with /l/ or /r/ (e.g., rock-lock), opposed to /l/ and /r/ being trained word initially and tested in all environments[11]. Logan et al. found training could generalise to natural speech samples, with results indicating a moderate, yet significant improvement from participants' pretest to posttest scores.

Since Logan and his colleagues published their first HVPT studies, numerous L2 speech researchers have employed HVPT to great effect (e.g., Saito, 2018; Thomson, 2018), finding more talkers preferable to fewer talkers and natural speech samples preferable to synthetic speech samples (Ingvalson et al., 2014)[12]. The utility of HVPT has been demonstrated with both segmental and suprasegmental contrasts, and in a wide array of languages, evidenced by research in French vowels (Brosseau-Lapre et al., 2013), Hindi stop consonants (Pruitt et al., 2006), Hungarian vowels (Archila-Suerte et al., 2016), Japanese vowel length (Hirata et al., 2007), and Mandarin tones (Lee and Hwang, 2016).

Despite its name, HVPT's "high variability" is a subjective if not misleading label as variability typically refers to the use of different talkers rather than phonological environment (e.g., diverse words or connected speech). Thomson (2018) conducted a review of 32 HVPT studies, finding that

[11] Strange and Dittmann used different methods for testing and training. They trained listeners on a same-different task and tested using an oddity task. Further, they trained participants using /l/ and /r/ word initially, but tested /l/ and /r/ perception in all environments, including initially (e.g., rock-lock), intervocalically (e.g., berry-belly), in consonant clusters (e.g., broom-bloom), and word finally (e.g., core-coal).

[12] Enhancements in artificial intelligence have evolved and improved since the Ingvalson et al., 2014. Today's (post-2015) synthesis generates spectrograms and exploits manipulated utterances enhanced through AI. In the near future, it will be possible to realistically manipulate speech samples to the extent that it is indetectable to the human ear; however, at present, it remains challenging to effectively employ speech synthesis systems.

while generalisation tasks (i.e., the assessment component of HVPT training) can implement new words, studies which include them are “less common” (p. 215).

In a well cited study, Thomson (2012) used HVPT to train native Mandarin speakers to identify 10 Canadian English vowels (/i, ɪ, e, ɛ, æ, ɒ, ʌ, o, ʊ, u/). The training couched target vowels in the consonantal frames /pV/ and /bV/, produced by a male talker. To avoid potential orthographical confounds, Thomson used distinctive nautical flags to represent each sound. When participants heard a given syllable, they were to identify it by mouse-clicking the appropriate flag in the computer program, and immediate feedback was provided (an audible “chirp” for a correct answer, and a “beep” plus a flashing image of the correct flag for an incorrect answer). Training consisted of eight sessions over three weeks. Thomson conducted two tests to explore how well the training might generalise to broader contexts. The testing contexts included /pV/, /bV/, /gV/, /kV/, /zV/, and /sV/ syllables produced by a female talker. In addition to indicating which syllable the participants heard, they were also asked to indicate how confident they were in identifying the correct sound by using an 8-point scale. The stimuli were randomly presented and modest, yet significant improvements were made in the average scores between time 1 and time 2, as well as in noted confidence in selecting the correct sound. It would have been interesting to see how well the training would generalise to vowel perception in English words or whether gains would be limited to laboratory-suited syllables.

Other studies, such as Iverson et al. (2012), have extended variability to diverse words in training, but not testing. In Iverson et al., a traditional /bVt/ frame for pre- and post-testing was employed. It would seem appropriate to have tested how an experimental phonological environment (perhaps a few) might generalise outward to other environments rather than the opposite; however, regardless of the method, results were positive, showing that training on a variety on environments improves perception in a single environment. Such studies show that diverse words can be useful for training.

As a caveat, individual differences place limitations on which groups variable stimuli (listening prompts) are suitable for. For instance, perceptual limitations of lower-aptitude learners (Perrachione et al. 2011) and children (Brekelmans et al. 2020) lead to poor performance with high variability and better performance with low variability. A more appropriate group, therefore, would be an adult cohort with more advanced language skills (or above a level which may be considered "low-aptitude").

Summarising HVPT, through its use of assessment, HVPT is a natural outlet for the present research. This research will use multiple talkers and high-intermediate to advanced listeners, compatible with HVPT use and literature.

Having established the type of prompts used to explore vowel perception, additional, theoretical considerations explore the suitability of prompts for assessing vowel perception. The remainder of the literature review discusses the attributes of authenticity and validity.

Authenticity. Authenticity has garnered the attention of scholars in pronunciation (Flege et al., 1996; Isaacs, 2014), listening (Field, 2019; Wagner, 2021) and language assessment (Bachman, 1990; Bachman & Palmer, 1996) for decades, and inching toward it constitutes a motivating factor for the current research. The current research has been informed by the domains of language testing, L2 listening and speech perception, and each were referenced to create a working definition of authenticity. (A more complete cross-domain overview of the varied definitions of authenticity can be found in Gilmore, 2019.)

Bachman and Palmer (1996), cited regularly in test development contexts (e.g., Field, 2019; Ockey & Wagner, 2018), explain authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a [target language use] task" (p. 23). Their emphasis is on *tasks* and their *correspondence* to a *specific* real-world situation. For listening assessment, Field (2019) proposes *cognitive processes*, as advocated by Weir (2005) to be included when defining authenticity. Authenticity is not typically defined in speech perception research or discussed in related research design texts, but the terms has been used to refer to speech characteristics which are native-like, such as voice onset time (Flege, 1991) and rhythm (Dickerson, 2016), as well as non-native-like, such as authentic Chinese accented English (McGowan, 2015). In such studies, there is a one-to-one correspondence between the speaker and authenticity, presupposing the recording and use of *human speech*. For the purpose of the thesis, then, authenticity is expressed as a reflection of tasks, materials, and cognitive processes which have been produced or elicited by human speech. Prompts will thus be designed to reflect this purpose. Where newly designed prompt types differ from canonically employed prompt types, the effect on performance will be explicated.

This study does not claim or endeavour to employ materials which are labelled "authentic" as a truly authentic assessment is not likely to be fruitful (Bachman, 1990). There may be a gap between competence and performance (Everington et al., 2007), as well as a gap between performance and score interpretation (Messick, 1996). Consequently, there is a paradoxical limitation for those striving for authenticity: what is considered locally authentic to a given assessment or scholarly work is inherently globally inauthentic to the real world. As Messick (1996) states, "ideal forms of authenticity and directness rarely if ever exist"; compromising between authentic and inauthentic conditions is not only practical, but essential.

The binary view of authentic or inauthentic may not be useful as each individual will have a subjective perspective of what it constitutes. Instead, it may be beneficial to bypass the authentic-inauthentic dichotomy by considering authenticity as a continuum (Pinner, 2014; Wagner, 2014). Rather than flipping a metaphorical switch and transposing inauthentic with authentic, the current research aims at moving closer toward the spectral pole of authenticity than is typically permitted in laboratory studies of listening perception research. Conscious decisions were made to include authentic elements to tasks, materials, and processes where possible, including the use of real words and locations, naturalistic speech (i.e., human generated rather than synthetic), and sentences where bottom-up and top-down processes may be engaged for assessing vowel perception.

Validity. Are the claims being made justified by the evidence possessed? This epistemological inquiry is the crux of assessment validity (Bachman & Palmer, 2010; Kane 2016; Knoch & Chapelle, 2017). Validity, used here, is the justifiability of using an assessment for a given purpose. Two threats to validity are construct irrelevant variance and construct underrepresentation (Messick, 1996). Construct irrelevant variance occurs when there are variances in observed performance which may be attributed to something other than the skill being assessed. This is equivalent to confounding variables undermining the results of an experiment. Implicitly, construct irrelevant variance is what modern speech perception researchers are combatting by reducing stimuli to fixed consonantal frames (e.g., bVt, bVbə, hVd). Just as important, however, is how representative the stimuli are to the skill or trait (e.g., listening proficiency) being explored. If not enough of the target skill is being assessed to support claims or inferences being made about the trait, validity is threatened by construct underrepresentation.

Construct underrepresentation is akin to judging a pilot's aviation skills on based on landing a single virtual plane on a single runway in a flight simulator. Landing is an important factor in flying a plane, yet landing reflects only one component of piloting. Further, doing so on a single runway may not reliably translate to landing another runway. Making accurate decisions on the prospective pilot's overall proficiency based on such restrictive information would be nearly incidental, especially when generalising from an artificial environment. Though an excellent pilot would be expected to do well with the landing, a pilot who is far less adept may also do well at landing—there are automated means, for instance, which augment the pilot's skill. Similarly for vowel perception, when assessing a learner's ability to identify or distinguish between L2 vowels, restrictive prompts types may not necessarily provide sufficient information to make judgements about the learner's actual ability. The present research, therefore, introduces more varied environments for assessment and explores the extent to which restrictive contexts generalise to additional contexts.

A degree of construct underrepresentation and overrepresentation is inevitable (Messick, 1996). Language competence cannot be directly viewed or measured. Beyond specific, contrived contexts, it is unlikely—if not impossible—to develop a measure that fully reflects a given set of language skills. Consequently, the goal is not to eliminate all conceivable avenues of construct underrepresentation, but to eliminate any which would prevent competence to be demonstrated, or that would allow a lack of competence to be demonstrated so that effective, systematic training may take place. In this way, a valuable aim is to identify listening prompts which may best reflect the construct of L2 vowel perception. Should an isolated, phonologically restrictive context generalise well to other more varied contexts, that would reflect a well-constructed prompt type for its purpose, regardless of its limited phonological constitution.

There is a balance between the threats to validity: too little of what the researcher is trying to assess (construct underrepresentation) and results may fail to generalise (Harding, 2017); too much in what the researcher is trying to assess (construct irrelevant variance) and there is a risk not being able to properly interpret results as they may not be due to manipulation of independent variable. As can be surmised by the strict adherence to fixed consonantal frames (e.g., monosyllabic words in hVd or bVd frames) and the lack of variety in speech perception stimuli in published research, speech perception testing has tacitly endorsed construct underrepresentation over construct irrelevant variance.

The question remains, are the claims a researcher would like to make about a learner's ability to differentiate between target vowels justified by the evidence provided by the assessment? To answer this, it is important to ensure the construct is sufficiently represented by the prompts employed to measure it. If diverse stimuli perform the same as or similarly to fixed stimuli, and if diverse stimuli offer few systematic differences in performance, it is reasonable (or valid) to conclude that the evidence justifies the claims. If participants perform significantly differently with diverse listening prompts, and if types of prompts can be mapped to difficulty levels, the answer is that many of our claims of what does and does not work for assessing L2 speech perception will need revisiting.

Addressing intelligibility and accuracy

Before concluding the literature review, a brief mention of intelligibility and accuracy will help orient the research. Intelligibility over accuracy has been identified as a key aim of L2 pronunciation research (Isaacs & Harding, 2017; Kang et al., 2020; Thomson, 2017), with the predominant operational definition offered by Derwing and Munro (1997, 2005). The prevailing position is that accent—how a speaker's utterance approximates a target language or dialect—is separate from the ability to be understood (intelligibility) and the ease at which it is understood (comprehensibility). A speaker may have an accent and still be intelligible and comprehensible, making accuracy a tertiary concern after intelligibility and comprehensibility.

How intelligibility and comprehensibility interact is succinctly illustrated in a matrix (Table 1) extracted from Munro and Derwing (2015, p. 380), with low intelligibility and high comprehensibility of key interest for the present research.

Table 1. Results of possible intelligibility and comprehensibility combinations

Intelligibility	Comprehensibility	Result
High	High	Utterance is fully understood; little effort required
High	Low	Utterance is fully understood; great effort is required
Low	Low	Utterance is not (fully) understood; great effort is exerted
Low	High	Probably rare. Utterance is not fully understood; however, the listener has the false impression of having easily determined the speaker's intended meaning

While it may be deemed "probably rare" that two interlocutors expect a message has been successfully communicated while an entirely different message has been understood by both parties, it remains a predictable (based on our understanding of typically conflated L2 segments) and non-trivial concern, constituting a glaring theoretical blind spot.

For ambiguous contexts, where redundancy (e.g., syntax or context) does not enable a listener to fill-in incomplete information (such as context filling the gap created by an L2 learner's poor vowel perception), accuracy and intelligibility may presumably intersect. The ability to understand the speaker's message, for instance, may be contingent upon the listener's ability to accurately differentiate between two L2 vowels. This can be illustrated with Grenville Street and Granville Street, London streets less than a mile apart. If two individuals agree to meet at Granville Street using spoken communication, and the listener is an L2 learner who cannot reliably differentiate between /ɛ/ and /æ/, the listener may end up at Grenville Street, as described in Table 1. Even if the speaker were aware that both Grenville Street and Granville Street exist, the listener's difficulty in differentiating between specific vowels may not be apparent. The context and syntax in such a situation cannot resolve the listener's tendency to conflate the L2 vowels.

To reflect such ambiguity, the present research will explore the convergence of intelligibility and accuracy, employing listening prompts which are syntactically and situationally equivocal.

Pilot

Research aims and questions for the pilot

The purpose of the present doctoral research was to investigate the use of diverse prompt types—and ultimately connected speech—for assessing vowel perception in English second language (L2) learners. Commencing this exploration, a pilot was conducted to help assess the potential use of connected speech in an oddity discrimination paradigm, identify suitable items for the primary study, and uncover operational areas for improvement before embarking on the principal research.

Questions the pilot attempted to answer included:

1. To what extent is the four-interval (4I) oddity task suitable (fit for purpose) for Sentences in the primary study?
2. To what extent do administration sequence, syllable count, vowel type, word frequency, and talker influence participant performance and how might this inform the primary study?
3. To what extent is the study's operational design (platform) suitable for the primary study?

Research design

The pilot addressed questions with a predominantly quantitative design, but included a short, informal questioning immediately after participants had completed the experiment. Participant questioning helped inform decisions about the efficacy of the 4I-oddity task and prompt types by including a participant perspective. The remainder of the pilot was analysed quantitatively. Specific analyses to address each question are provided in the section, Data Analysis.

Methodology

Participants

Talkers. Five voice actors (3 male, 2 female) from the South of England were hired to produce recordings for the experiment. Actors, opposed to university student volunteers, were chosen for practical purposes: they were familiar with voice recording, were able to record and re-record readily, and enabled commercial rights to be obtained in exchange for remuneration. Actors were recruited through Upwork.com, an online freelancing platform. Voice actors ranged in age from 22-55 years old. The first talker was employed to identify potential issues and improvements relating to the provided word lists, but the talker's recordings were not used for the experiment, leaving audio recordings from a balanced proportion of 2 male and 2 female talkers.

The chosen English L1 dialect was employed for practicality. The single English variety was reasoned to promote familiarity for the listener (Clopper, 2021; Njie et al., 2022) and the South of England variety reflected the university's geographic location. I acknowledge that selecting a given accent or variety has socio-linguistic implications, and asserts that the appropriate language and variety for assessment, research and training purposes depends on learner needs and what is required for a given context.

Listeners. Ten adult London-based postgraduate university students from Mandarin (n = 7), Korean (n = 2), and Japanese L1 (n = 1) L1 backgrounds were recruited using direct recruitment and snowball sampling. All participants, who reported having normal hearing, had received a minimum overall IELTS score of 6.5 and no score of less than 6.0 on listening, speaking, reading and writing or passed the university's pre-sessional exit exam. Listeners were not financially compensated for their time, but received immediate feedback on their performance, which may have helped identify areas of relative strength or areas for potential improvement.

Instrumentation

Language background questionnaire. Participants completed a paper-based language background questionnaire prior to the online portion of the experiment. The questionnaire was based on Jones (2015) and modified to include information specific to English language educators, a student cohort I (the researcher) had access to during the recruitment phase of Experiment 1. The questionnaire identified participants' English language exposure, L1, age, gender, and teaching experience for potential groupings during analysis.

Perception experiment design

Discrimination task. The pilot employed a four-interval (tetrad) oddity design (Mitterer & Mattys, 2016; Rogers, 2017) to examine discrimination. Listeners heard four isolated words in sequence, with one word distinguishable from the others by the presence of a contrastive vowel phoneme (e.g., leave-live-leave-leave), and had to indicate which of the four words was the different word. Each token was spoken by a different talker (Brekelmans et al., 2020, Flege et al., 1994). The talker place in the sequence, place of correct answers, and number of correct answers per talker were proportionally distributed. Half of the tetrads contained words differentiated by the /i, ɪ/ contrast, the other half were differentiated by /ɛ, æ/. After hearing the sequence of four words, participants indicated which word was semantically different than the other three by mouse-clicking a button on screen marked 1, 2, 3, or 4.

The oddity task was chosen over other common discrimination tasks, such as same-different (AX), to combat response bias (McGuire, 2010) and chance of correctly guessing (Schouten et al., 2003), and for ease of instruction during the experiment. At the risk of additional cognitive load (Mitterer and Mattys, 2017), four rather than three intervals was implemented to reduce correct answers derived by chance. As the study employed a roving design (the odd word was found in any of the four possible positions) and words beyond bVd were used, it was uncertain whether the fourth option would be detrimental to participant performance. Helping counteract effects of cognitive load, the study incorporated a replay option (McGuire, 2010). That, combined with the use of higher level English language speakers for this study, was hoped to alleviate cognitive load concerns. The pilot was used to identify whether this design choice worked, or whether refinement was necessary.

Listening materials (stimuli). The stimuli were stressed target vowels embedded in isolated words and connected speech. Target vowels were /i, ɪ, ɛ, æ/, selected for their high functional load (Brown, 1988; Catford, 1987) and well-established tendency for conflation by L2 speakers, including those with Mandarin, Korean, and Japanese as their first language (Jones, 2015). Target vowels were embedded in three types of listening prompts: isolated bVd words (bead, bid, bed, bad), isolated words in diverse phonological environments, and connected speech. The bVd frame (Flege, 2021; Flege & MacKay, 2004) was used as a means of comparison for the more phonologically and sententially diverse prompt types. The “diverse” stimuli were created to provide a sample of the various environments which may colour target vowels (i.e., place, manner, and voicing, positioning before and after the target vowel, syllable shape).

Since the pilot was used to identify suitable prompts (items) for the primary experiment, a larger number of items was presented to participants than intended for the primary experiment. All minimal pairs were included regardless of frequency or potential for conflation (later used as selection criteria). This enabled a curated selection of prompts based on performance and other selection criteria. A complete list of prompts used for the pilot is found in Appendix: Pilot item lists[13].

The vowel pairs were presented together in three blocks of prompt types: bVd, Diverse Words, and Sentences. The bVd prompts contained 32 items for each vowel pair (16 per vowel). With four vowels, tokens for each talker were reused four times**,** once for each serial position. The Diverse Words held 115 items for /i, ɪ/ and 102 for /ɛ, æ/, and Sentences held 58 items for /i, ɪ/ and 52 for /ɛ, æ/. Diverse Words and Sentences were spoken by one of four talkers, meaning there was one iteration of each prompt, opposed to four. This enabled a larger number of prompts to be used for the exploratory study.

[13] Not all items that were recorded were used in the pilot.

Recordings. Listening materials were read aloud and recorded by four voice actors (see Talkers) in acoustically attenuated sound booths. Settings were 16 bit depth at a sampling rate of 44.1 kHz. Recordings were made in different studios and were subsequently matched for volume. WAV files were matched in Adobe Audition CC using the International Telecommunication Union Radiocommunication broadcast standard ITU-R BS.1770-3 (Lee et al., 2015). Talkers were instructed to say isolated words in the carrier sentence, “I said [target word]”, and to speak at approximately 150 words per minute, a speech rate congruent with the range required for accurate L2 comprehension (Griffiths, 1992). For words with common alternative pronunciations, Talkers were directed toward the desirable pronunciation for the experiment. For instance, Whitfield may be pronounced by some speakers as /hwɪtfild/[14], with an /h/ preceding the /w/, while others may pronounce the word as /wɪtfild/, with no word-initial /h/. Here, talkers were directed to pronounce Whitfield (and other similar “wh” words) without an initial /h/. Talkers spoke each utterance three times, with the clearest production and recording being selected for use, except for bVt utterances. The bVt utterances were each used to create unique bVt sequences for the bVt block. Words and sentences were re-recorded as needed. One set of talker recordings was replaced after the initial pilot as the audio (speech rate, sound quality) was inconsistent with the other three talkers. Prior to upload, WAV files were converted to MP3 file format (128 kbps) for compatibility with the online experiment platform.

Platform. The pilot listening experiment was conducted using TP 3.1 (Rauber et al., 2013), a freeware application for conducting speech perception experiments, on a laptop computer. The platform was selected for its accessibility and ease of use in constructing oddity tasks. The platform was accompanied by a physical printout of the background questionnaire.

Procedure

Participants completed a language background questionnaire (see Appendix) just prior to the listening perception experiment. The perception experiment was administered in a quiet room using over-ear headphones and a laptop computer. Instructions were explained to each participant orally. After instructions were provided, participants clicked “Start” in the programme, initiating a block of seven practice trials before the formal experiment was presented. The practice block was not used for statistical analysis. For each section of the experiment, the screen displayed a similar user interface, with four clickable buttons labelled 1, 2, 3, and 4, a replay button in the bottom left of the window, and an exit button on the lower right. The section heading (e.g., bVd words) was present at the top of the window, and progress (current item number and total item number) was at the top right. The first block of listening prompts consisted of bVd words to provide a baseline of performance. After bVd, one group (*n* = 5) received isolated diverse words prompts followed by a block of connected speech prompts, while a second group (*n* = 5) received a block of connected speech prompts followed by isolated diverse words prompts.

Results were provided in a pop-up screen after the listening components had been finished, displaying the aggregate account of correct and incorrect responses. Upon completion, participants took part in a short, unstructured interview about perceived item type difficulty, preferences, and any memory challenges associated with recalling prompts[15] (e.g., “Was using four utterances in a trial too challenging to remember?”).

Data Analysis

Suitability of the 4I-oddity task for sentences was determined by considering results from internal consistency (Cronbach’s alpha), an item analysis, the post-experiment interviews with participants, and the needs of the primary study. These elements are reported separately in Results and converged in Discussion.

[14] The broad /hw/ transcription was elected for common understanding, and may be represented as a single IPA grapheme, “ʍ”, resulting in /ʍɪtfild/.

[15] Four rather than three options were used in the pilot to minimise the effects of guessing, but with sentences, remained a salient concern due to cognitive load.

Internal consistency, calculated by Cronbach's alpha, has been cited as an integral tool for instrument development (Cortina, 1993) and remains prevalent in the literature for identifying the reliability of assessment measures (Hoekstra et al., 2019).

The item analysis helped identify how well items and prompt types were performing. Item analysis includes indices for difficulty (the proportion of correct responses, or *Pc*) and discrimination[16] (*Dc*), an estimate of how well an item differentiates high and low performing participants. Item analysis is often done for educational assessments using large numbers of participants, with a minimum sample of 100 suggested (Nevo, 1980; Park, 2019). This was not possible for the pilot, nor desirable given the pilot's low stakes nature. Results were thus intended to guide design decisions in conjunction with other factors, such as phonological context (e.g., neighbouring consonants) and word frequency, rather than constrain item choices with specific decision criteria. Overly difficult items ($Pc < .5$) were less desirable, as were negatively discriminating items. Negative discrimination indicates lower performers on the assessment did well on an item, while higher performers on the assessment did poorly on the item. Indices were averaged to identify the broader functioning of each prompt type (Dudycha & Carpenter, 1973). The Sentences prompt type was expected to be more difficult than the isolated prompts as there was no indication which segmental feature in the sentence made the odd sentence odd, and there was considerably higher cognitive load listening for the difference among four sentences compared to four isolated words.

Prompt type performance was also examined by comparing the bVd words with Diverse Words and Sentences. A repeated measures analysis of variance (ANOVA) was conducted to examine the significance of any differences in mean performance as a function of. prompt type.

Potential effects of administration sequence, syllable count, vowel type, word frequency, and talker were investigated. There were two administrations sequences in the present study, explored with an independent samples *t*-test. To explore the effect of syllable count, mono- and multisyllabic word performances were converted to proportion correct, then compared in a

[16] Discrimination (point biserial correlation) is a measure of how well the item correlates with total score.

dependent samples *t*-test. Likewise, vowel type was explored through proportion correct and a dependent samples *t*-test. Word frequency was split into two analyses: frequency and difference in frequency (DF). Frequency was determined using the British National Corpus (explained in depth in Diverse Words (isolated words). To obtain the difference in frequency, the frequency of the less frequent word was subtracted from the frequency of the more frequent word. For instance, the frequency total for “fill”, 10925, was subtracted from the frequency total of “feel”, 42379, to get a difference in frequency of 31454. Finally, the effect of talker was investigated with a repeated measures ANOVA.

Operational aspects of the study’s design focused on the platform from an administrative perspective, including its ability to perform as intended and its scalability. Unlike previous questions which had quantitative measures to guide decision making, these are subjective considerations based on researcher judgment.

Results

Suitability of using sentences in a 4I oddity task (compared to bVd and Diverse Words).

Identifying suitability of the prompt type began with Cronbach’s alpha. The experiment was internally consistent in all contexts, though Sentential prompts was least of the three prompt types. Cronbach’s alpha for /i, ɪ / bVd, Diverse Words, and Sentences was .93, .95, and .82, respectively. For /ɛ, æ/, Cronbach’s alpha was .89, .93, and .80 for bVd, Diverse Words, and Sentences, respectively, revealing strong internal consistency suitable for a research experiment (Taber, 2017).

Next, item performance was conducted. Items generally performed well and offered suitable options to select from for Experiment 1 (see Appendix Pilot item analysis report. Mean performance indices by prompt type are displayed in **Table** .

Table 2. Pilot mean difficulty (proportion correct) and discrimination (point biserial correlation) indices for bVd, Diverse Words, and Sentences by vowel pair

	/i, ɪ/		/ɛ, æ/	
Prompt type	Difficulty	Discrimination	Difficulty	Discrimination

bVd	.73 (.24)	.50 (.30)	.71(.21)	.40(.25)
Diverse Words	.76 (.16)	.43 (.36)	.67(.15)	.37(.29)
Sentences	.51 (.14)	.26 (.37)	.46(.14)	.25(.32)

Note. Standard deviations are in parentheses.

As seen in Table , the bVd and Diverse Words prompt types performed similarly for both vowel pairs, with comparatively easy, yet effectively discriminating prompts (*Dc* > .2). Expectedly, Sentences was the most challenging of the prompts. The discrimination indices for Sentences were a harbinger, however, as standard deviations were larger than the means, indicating negative values were within one standard deviation. A negative discrimination value occurs when poor performing participants (overall) perform well on an item, while high performing participants (overall) perform poorly on an item. The irregular performance for Sentences was a red flag for the prompt type. Memory confounds were considered to be a contributing factor, causing otherwise strong participants to do poorly, while relatively poorer performing participants who had stronger short-term phonological memory had an advantage.

A repeated measures ANOVA found a significant effect for context in both vowel pairs (/i, ɪ/, $F(2, 18) = 9.492$, $p = .002$, $\eta^2 = .51$, and in /ɛ, æ/, $F(2, 18) = 22.754$, $p < .001$, $\eta^2 = .73$). Eta squared suggested that 51% and 73% of the variance in scores, for /i, ɪ/ and /ɛ, æ/ respectively, was due to the main effect, context. To confirm where the significant differences were, a post hoc pairwise comparison with Bonferroni correction was conducted. For both vowel pairs, differences were significant between bVd and Sentences ($p < .05$) and between Diverse Words and Sentences ($p < .05$), but not between bVd and Diverse Words ($p > .05$).[17]

[17] Pearson correlations were run to identify the association between performance with isolated prompts and performance with Sentences. The /i, ɪ/ bVd contrast was not significantly related to Diverse Words or Sentences ($p > .05$), while Diverse Words was positively correlated with Sentences, $r(9) = .70$, $p = .02$. For /ɛ, æ/, bVd was significantly positively associated with both Diverse Words ($r(9) = .73$, $p = .02$) and Sentences ($r(9) = .70$, $p = .03$), though the relationship was strongest between Diverse Words and Sentences ($r(9) = .85$, $p < .01$). It appears Diverse Words may be a better predictor than the fixed frame prompt, but the predictive utility of isolated prompts for performance in sentences will need to be further explored with a larger sample in Experiment 1.

Variable effects. Five potential effects were considered: administration sequence, syllable count, vowel type, word frequency, and talker. Each are sections separately below.

Sequence. An independent samples *t*-test was conducted to check for possible sequence effects during administration. The 10 participants were split into two groups of five. Group 1 took the sequence bVd-Diverse Words-Sentences, obtaining mean scores of 162 (*SD* = 17.7) and 52.4 (*SD* = 16.8) for Diverse Words and Sentences, respectively. Group 2 was administered the sequence bVd-Sentences-Diverse Words, obtaining mean scores of 148 (*SD* = 40.3) and 51.8 (*SD* =11.4) for Diverse Words and Sentence, respectively. Independent samples *t*-tests between the groups revealed the differences were not statistically significant (*p* > .05). If the present design is used for the primary study, sequence is not expected to be a biasing factor.

Syllable count. Exploring the possible effect of syllable count, Diverse Words[18] items were first split into mono- and multisyllabic groups. For /i, ɪ/, monosyllabic words (*M* = .78, *SD* = .14) had a higher proportion correct than multisyllabic (*M* = .71, *SD* = 23); however, the difference was not found to be significant (*p* > .05). For /ɛ, æ/, the difference was greater as monosyllabic words (*M* = .79, *SD* = .12) had a higher proportion correct than multisyllabic (*M* = .58, *SD* = 19). The difference was significant (*t*(9) = 7.514, *p* < .001).

A controlled set of target vowels with shared neighbouring segments (e.g., band-banding, bend-bending) was then isolated. This controlled set included 18 monosyllabic, 20 disyllabic, and 2 trisyllabic words, as shown in Appendix. One monosyllabic word (pat) was excluded as it did not properly play for participants during several test sessions, thus eliminating the pat-patter pair from the set and leaving 17 monosyllabic and disyllabic word pairs for comparison, plus the initial 2 disyllabic and trisyllabic word pairs.

Overall, scores for the controlled set were Monosyllabic (*M* = .72; *SD* = .12), Disyllabic (*M* = .60; *SD* = .16), and Trisyllabic (*M* = .47; *SD* = .15). A repeated measures ANOVA was conducted and

[18] Diverse Words were used exclusively as bVd Words lacked contrasts while Sentences offered too few cases for comparison.

Mauchly's test of sphericity was significant ($p < .05$), leading to the use of a Greenhouse-Geisser correction. Results indicated an effect for syllabicity $F(2, 18) = 26.958$, $p < .001$, $\eta^2 = .750$. Eta squared suggests that when phonological context was controlled, 84% of variability in our pilot scores was accounted for by syllabicity. Effects of syllabicity should be explored in the primary study.

Vowel type. Vowel type was next investigated. For /i, ɪ,/, /ɪ/ (*M* = .78, *SD* = 14) was identified at a non-significantly ($p > .05$) higher rate than /i/ (*M* = .74, *SD* = 18). For /ɛ, æ/, there was a significant difference in scores between the /æ/ (*M* = .73, *SD* = 14) and /ɛ/ (*M* = .62, *SD* = .17) vowel pair ($t(9) = 3.778$, $p < .01$). Looking at L1 phonology, the Mandarin and Korean groups (n = 9) have an approximate equivalent for /i/ and /ɛ/ in their L1 phonological inventories, but not /ɪ/ or /æ/ (see Experiment 1, Performance predictions and PAM-L2.). Though a small sample, it appears vowels which had no direct L1 category to assimilate to (/ɪ/, /æ/) tended to be easier for participants to identify as "odd" than vowels which had an equivalent L1 category (/i/, /ɛ/).

A general pattern thus presented itself, though not always statistically significantly: single syllable words make identifying the odd word out easier than multi-syllabic words, and it may be easier for L2 listeners to identify odd utterances when the odd utterance is contrasted by a vowel which is absent from the listeners' L1. The two were then combined, with mean results found in **Table** .

Table 3. Matrix of syllable pair type by L1 vowel status showing proportion correct

L1 vowel status	Monosyllabic	Multisyllabic
No L1 equivalent (/ɪ/, /æ/)	.85(11)	.64(21)
L1 equivalent (/i/, /ɛ/)	.72(16)	.59(18)

Note. Standard deviations are in parentheses.

A repeated measures ANOVA violated the assumption of sphericity and a Greenhouse-Geisser correction was made. A significant difference was found between group means, $F(3, 27) = 11.32$, $p < .001$, $\eta^2 = .56$. Pairwise comparisons with a Bonferroni correction found a statistically significant ($p < .05$) difference between shorter words with /ɪ/ or /æ/ as the odd one out and longer words with /ɪ/ or /ɛ/ as the odd one out. A significant difference was also found between shorter words which had /ɪ/ or /ɛ/ as the odd one out and shorter words which had /æ/ or /ɛ/ as the odd one out.

It was not always the case that words with shorter syllables were perceived[19] at a higher rate than their longer counterparts, nor was it invariably the case that /ɪ/ was perceived at a higher rate than /i/ as the odd utterance, or that /æ/ was always correctly perceived as odd more readily than /ɛ/. Yet it was nearly always the case that a shorter utterance where the odd word contained /ɪ/ or /æ/ as the target vowel was perceived at a higher rate than longer utterances with /i/ or /ɛ/. This is readily visualised through a series of minimal syllable pair matrices using the controlled word set (individual matrices are shown in Individual vowel and syllable matrices).

[19] Perception is inferred by observed performance.

Word frequency. Absolute and relative word frequency effects were probed using linear regression. First, using data from Diverse Words, the proportion of correct responses per item was paired with word frequency. No significant relationship was found between word frequency and proportion correct ($p > .05$). Words which were common could be discriminated more, less, or equally accurately compared to words which were uncommon. A relative measure of frequency was then checked, converting frequency to a rank, a hierarchy of words in the word list based on frequency. As with absolute frequency, no relationship was found ($p > .05$). Lower ranked words, such as "scrim" (*Pc* = 1, rank = 97) sometimes held higher proportion correct (*Pc*) than higher ranked words, such as "lead" (*Pc* = .4, rank = 2); however the opposite was equally true, where a higher ranked word, such as "leave" (*Pc* = .9, rank = 4) held a higher P value than a lower ranked word, such as "tinny" (*Pc* = .3, rank = 93). As a final examination of frequency, the difference between frequencies of minimally paired words was checked. Again, no relationship was found ($p > .05$).

Talker. Overall participant scores (proportion correct) were divided by talker. Participant performance with Talker 1 (*M* = .62, *SD* = .15), Talker 2 (M= .61, *SD* = .11), and Talker 3 (*M* = .61, *SD* = .68) were similar to each other, but descriptively distinct from Talker 4 (*M* = .68, *SD* = .09). A repeated measures ANOVA (sphericity assumed, $p > .05$) displayed a significant global effect for Talker; $F(3, 27) = 4.163$, $p < .01$, $\eta^2 = .40$. Pairwise comparisons were run with a Bonferroni correction, showing the significant difference was between Talker 4 and all other talkers. As surmised by the descriptive statistics, Talkers 1-3 were not statistically different.

Consequent to this analysis, Talker 4's audio productions were revisited, revealing two issues which separated the talker from the others: hyper-articulation and the microphone used to record. Talkers 1-3 used condenser microphones, while Talker 4 used a dynamic microphone. Replacement was deemed necessary for the main study, and an additional talker was hired to re-record Talker 4's audio.

Suitability of the platform.

The platform was not found to be suitable. In general, several glitches were encountered which were not fully understood. Four items did not consistently play properly. Not only did this result in four items being removed from analysis, it also added time to the exam as it paused the experiment and caused confusion with affected participants. Pauses that were programmed into the platform to permit participants to take a break did not function properly. To combat this, participants were advised to avoid clicking “next” if they needed a break. However, two participants continued through the experiment and afterward lamented the lack of a built-in pauses. Providing an explicit option for pausing the test will be helpful in future iterations of the platform. The platform was also determined to be non-scalable. It cannot be readily added to university computers, and should that problem be resolved, results would need to be obtained manually and individually, meaning data from each participant will be on a separate file. The exported CSV files further had to be manipulated, and with a sample larger than 10, the process would be time intensive. Though the added time to format each participant would not be prohibitive in a larger sample size, the multitude of problems encountered with the platform make it a poor choice for the primary study. Consequently, a new platform will be employed.

Informal post-experiment questioning. Upon completion of the bVd, Diverse Words, and Sentence tasks, participants were asked about their experience. Key questions regarded how participants felt about four options and whether fewer would be better, and whether a particular section was preferred over others.

Participant opinions about 4I oddity. Several participants used the term “fine” to describe choosing between four options opposed to fewer, with one stating that being able to concentrate on distinguishing between four options was “not a problem”. One participant, noting that Sentences was the most difficult section, mentioned they “felt like there were always two wrong and two right”. For this test taker, it appears three options would have made a process of elimination (i.e., guessing) considerably easier. As four options was a consensus “fine” and it lessens the chances of guessing the correct answer, the pilot has provided support for retaining this test feature.

Participant opinions about prompt types. Responses were mixed. Sections were seen by many participants as different, but not better or worse. One participant noted that they were tired during Sentences. This was the highest performing participant, with an overall score of 84%. The participant's scores did drop in the Sentence section to 64% and 74% for /i, ɪ/, and /ɛ, æ/; however, scores for all participants were on average 33% lower in Sentences than in other sections. This participant was part of Group 1, meaning Sentences was administered last. Jane, from Group 2, noted being "fatigued a little", and that Diverse Words were easier than Sentences. Jane also stated, "When I listened to Sentences, I couldn't tell the difference". This was a popular sentiment and explicitly noted by three participants. Jane further mentioned that she preferred Diverse Words to Sentences. Brad felt bVd was the most difficult, but couldn't explain why. Alice performed strongly with bVd, but poorly in Sentences, and was asked, "Do you think bVd accurately represented your current ability to discriminate between these vowel sounds?", to which she responded, "No, you've got to do well with sentences. We don't speak with individual words". She then explained, "Sentences were the most difficult. I felt like there were always two wrong and two right. Especially when you're not familiar with the word, like Fenn Street and Fann Street". Ironically, this participant scored better with Fenn/Fann Street in Sentences than most other words. The mixed results would suggest that the selection of one section over another, or even including all in one test, should be up to the researcher's interests and participants' needs. How well each section addresses those needs will be explored during the primary study.

Pilot discussion

A pilot study was conducted to determine the suitability of using sentences in an oddity task for assessing vowel perception in Sentences, of individual items for use as prompts, and of the platform in general, and to investigate potential confounds to be mindful of in the primary experiment. Results were mixed for the Sentence prompt. The prompt type was adequately internally consistent and, in general, participants felt sentences with four options was challenging, but "fine". However, given the stark contrast in scores between Sentences and the isolated word prompts in conjunction with the relative prevalence of negative discrimination indices, the oddity task should be replaced with another. A forced choice identification task, for instance, would require far less cognitive load. The oddity task was chosen for the pilot because it offered no indication to the listener what distinction to listen for, it had no potential spelling confound (only having to select numerals 1-4), and minimised chance performance to 25%. That said, forced choice identification tasks would alleviate much of the cognitive load participants may have experienced, and are common in L2 speech perception studies (e.g., Balas, 2018; Iverson et al., 2012; Lengeris, 2009; Leong et al., 2018; Strange & Dittman, 1984). Replacing the Sentence oddity task with identification is seemingly the appropriate solution. The common use of fixed frame oddity and identification tasks makes their communal effectiveness in predicting sentence performance an emergent area of interest.

Individual item indices suggested general utility of the prompts. They will be considered further along with other measures such as word frequency (despite the lack of significance found in the pilot) and syllable number.

The platform was easy to use and populate with items, but resulted in too many errors and was not readily scalable for a larger study. It will be replaced with a more robust, reliable platform.

Finally, the pilot helped uncover several potential biases and areas of interest to be mindful of in the primary study. Syllable count will be investigated. It interacted with vowel type and may be an interesting predictor for performance. Word frequency should be investigated as part of “due diligence”, but expectations of significance tempered. Finally, the pilot showed listeners performed equally well with three of four talkers, but investigation into the fourth suggested a need for replacement with another voice actor.

Experiment 1

Methods

Participants

Talkers. Four British English voice actors with a South of England dialect were hired to produce recordings for the listening experiments. A change from the pilot talker group was made as one of the voice actors was replaced due to prompts performing statistically differently than the other three talkers.

Listeners. Forty-three adult students from a London, UK, university took part in Experiment 1 (age 18-41, *M* = 26.44 years, *SD* = 5.1). Thirty-eight participants were target L2 speakers (30 Mandarin, 8 Korean L1), 5 were a control. The L2 group began learning English at 3-19 years (*M* = 9.5, *SD* = 4.9), had studied English an average of 14.8 years (*SD* = 5.4), and had previously demonstrated their English language proficiency in a standardised test, with an average overall IELTS score of 7.3 (*SD* = 0.5) and IELTS listening subscore of 7.9 (*SD* = 0.7). Self-reported proficiency from scalar data (see Appendix X Language Background Questionnaire) showed participants identified as advanced (n = 10), high intermediate (n = 19), intermediate (n = 8), and low intermediate (n = 1) for overall English ability. For listening, participants indicated their ability as advanced (n = 11), high intermediate (n = 14), intermediate (n = 12), and low intermediate (n = 1). Mean age of arrival was 25.8 years (*SD* = 4.7). This group was predominantly recruited from an English teaching cohort. Twenty-eight participants reported having taught English for an average 1.6 years (*SD* = 2); 10 reported no English teaching experience.

Participants volunteered for the study through direct recruitment and snowball sampling; they received no monetary compensation for their participation. Included participants reported normal hearing as a pre-condition for participating in the study. One participant was excluded for reporting abnormal hearing.

The five-participant control—four English and one French-German L1—was predicted to readily distinguish between and identify target vowels based on the vowels being contrastive in their L1s. The French-German L1 speaker was an English trilingual with advanced education and training in phonetics and phonology. This control participant obtained a perfect score (30) on the TOEFL iBT Listening component, which converts to IELTS band 9 (ETS, 2021).

Instrumentation

Language Background Questionnaire. Participants completed a paper-based language background questionnaire, as described in the pilot, prior to the online portion of the experiment.

Listening Materials.

Listening materials (stimuli) were stressed vowels in isolated words and connected speech. Target vowel contrasts (/i, ɪ/, /ɛ, æ/) were selected for their high functional load (Brown, 1988, Catford, 1987) and established tendency for conflation by L2 speakers, including those from the target language groups (Jones, 2015). The target vowels were included in various phonological environments. Stimuli were grouped into blocks of discrimination and identification tasks (defined subsequently in this section) and presented as listening prompts to which participants would respond. Word lists and prompt types are described subsequently.

bVt Words (isolated words). The bVt words (i.e., beat, bit, bet, bat), were used for both discrimination and identification tasks. The bVt frame was chosen for its historically and current prominence in the literature (see A brief history of vowel perception). It was thus used as a benchmark to compare the performance of the more phonologically diverse prompts.

Diverse Words (isolated words). Diverse Words lists were created for comparison with canonical frames. Diversity included various phonological environments, syllables, and syllable structures in which to embed target vowels. The word lists contained eight words per target vowel (16 words per target pair). These words consisted of two canonical frames (bVd, hVd), two street names and four additional words. Canonical frames represented typical words employed in laboratory settings while street names presented actual London streets. Non-canonical words (including street names) were selected based on frequency and number of syllables (see Word and sentence list development.). The final word list for Diverse Words can be seen in Table 4 and Table .

Table 4. /i, ɪ/ Diverse Words frequency table

/i/ word	S-BNC Frequency	Grade level[1]	/ɪ/ word	S-BNC Frequency	Grade level[1]
Bead	7	K-5	bid	418	K-2
Heed	9	K-6	hid	36	K-2
Feel	3807	K-1	fill	641	K-1
Feeling	948	K-1	filling	171	K-1
Leave	3131	K-1	live	1888	K-1
Lever	44	K-3	liver	66	K-4
Sheep Lane[2]	2942	K-2	Ship Lane[2]	4558	K-2
Siemens Road[2]	275	n/a	Simmons Road[2]	339	n/a

Note. Frequency not used to determine inclusion for street names. Table shows raw frequencies from Spoken British National Corpus (S-BNC) and British National Corpus (BNC) grade level.

[1] Grade level was generated by the online platform, Compleat Lexical Tutor (Cobb, n.d.) based on the complete BNC corpus (100 million words) rather than the S-BNC (10 million words).

[2] Frequency given for the specific name (e.g., Simmons) without the generic (e.g., Road).

Table 5. /ɛ, æ/ Diverse Words frequency table

/ɛ/ Word	S-BNC frequency	Grade level[1]	/æ/ Word	S-BNC Frequency	Grade level[1]
bed	1839	K-1	bad	3076	K-1
head	1905	K-1	had	29316	K-1
bend	145	K-2	band	285	K-2
kettle	188	K-3	cattle	121	K-4
pet	114	K-3	pat	274	K-2
petter	n/a	n/a	patter	n/a	K-11
Fenn Street[2]	5	n/a	Fann/Fan Street[2]	106	n/a
Grenville Street[2]	3	n/a	Granville Street[2]	21	n/a

Note. Frequency not used to determine inclusion for street names (see Directions Sentences). Table shows raw frequencies from Spoken British National Corpus (S-BNC) and British National Corpus (BNC) grade level.

[1] Grade level was generated by the online platform, Compleat Lexical Tutor (Cobb, n.d.) based on the complete BNC corpus (100 million words) rather than the S-BNC (10 million words).

[2] Frequency given for the specific name (e.g., Simmons) without the generic (e.g., Road).

Directions (connected speech). Directions was an intermediate step between isolated word tasks and Diverse Sentences after considering physical equivalency (i.e., the carrier sentence is identical across prompts to ensure only the target word is examined). It is possible that the target word′s position in a sentence, sentence stress, or other physical alterations in prompts may compromise a listener's likelihood of correctly identifying a target word (Grant & Seitz, 2000). The Directions prompts therefore helped provide a comparison with Diverse Sentences. Controlling for physical equivalence could have been done by concatenating individual words together to make sentences (Grant & Seitz, 2000) or by excising target words from sentences (Pelzl et al., 2019). Neither would have resulted in naturally spoken speech. Concatenating words together would have resulted in disfluent sentences, while excising target words from diverse sentences to form the isolated words was not possible for all words in the final Diverse Words list (the diverse words outnumbered the number of unique sentences created). The Directions sentences provided more natural, fluent speech than would have been possible with concatenation or excision while implementing a more syntactically controlled environment and predictable context than found in Diverse Sentences.

The Directions sentence prompts extended the street names encountered in Diverse Words by including sentential context and adding five pairs of streets for the /ɛ, æ/ vowels. Streets were in carrier sentences which began with, "Meet me at", followed by the destination's street name (e.g., "Meet me at Ellen Street").

Inclusion criteria for destinations was being actual streets in London, displaying minimally paired specifics (e.g., Ellen-Allen) and having shared generics (e.g., Street). Frequency was considered problematic for street names and therefore not part of the inclusion criteria. Because included streets were authentic and local, participants were as likely to encounter a street name present in the corpus as not. There were 18 streets (nine pairs) in total. The full list of street names is found in Table .

Table 6. Street names used in Directions sentences

Target vowels	Cognate Street	Cognate Street
/i, ɪ/	Sheep Lane	Ship Lane
	Siemens/Seamen's Road	Simmons Road
/ɛ, æ/	Edison Road	Addison Road
	Ellen Street	Allen Street
	Epple Street	Apple Street
	Kemble Road	Campbell Road
	Fenn Street	Fann Street
	Grenville Street	Granville Street
	Redcliffe Square	Radcliffe Square

Note. Street names used with carrier sentence, "Meet me at…".

Diverse Sentences (connected speech). Diverse Sentences added syntactic and contextual diversity to the Directions sentences. Where Directions Sentences placed the target word in the penultimate position of the sentence, target words pairs could occur at any part of a sentence in Diverse Sentences. Unlike Directions, the context (topic) for Diverse Sentences was unpredictable.

Table 7. /i, ɪ/ Diverse Sentences target word frequency table

	S-BNC	Grade		S-BNC	Grade
/i/ Sentence	frequency	level[1]	/ɪ/ Sentence	frequency	level[1]
They conducted a faecal analysis.	0	K-12	They conducted a fickle analysis.	1	K-12
Feel the cavity first.	3807	K-1	Fill the cavity first.	641	K-1
Take the lead for me.	712	K-1	Take the lid for me.	107	K-3
The elderly man doesn't want to leave.	3131	K-1	The elderly man doesn't want to live.	1888	K-1
The Dutch have basic meals.	431	K-2	The Dutch have basic mills.	247	K-2
It was hard to see through the sleet.	10	K-8	It was hard to see through the slit.	11	K-6
It was a teeny audio file.	20	n/a	It was a tinny audio file.	0	K-2
The old man wheezed past me.	3	K-7	The old man whizzed past me.	16	K-4

Note. Table shows raw frequencies from Spoken British National Corpus (S-BNC) and British National Corpus (BNC) grade level. Frequency not used to determine street name inclusion (see Directions Sentences).

[1] Grade level was generated by the online platform, Compleat Lexical Tutor (Cobb, n.d.) based on the complete BNC corpus (100 million words) rather than the S-BNC (10 million words).

Table 8. / ɛ, æ/ Diverse Sentences target word frequency table

/ɛ/ Sentence	S-BNC frequency	Grade level[1]	/æ/ Sentence	S-BNC Frequency	Grade level[1]
Locate the efferent neuron.	0	K-19	Locate the afferent neuron.	0	K-17
Globalisation brought effluence to China.	6	n/a	Globalisation brought affluence to China.	6	K-10
Calculate the betting averages.	34	K-1	Calculate the batting averages.	10	K-3
I'd like to find a shop that sells gems.	29	K-4	I'd like to find a shop that sells jams.	169	K-2
I'd like a pedal board[2] for my birthday.	353	K-3	I'd like a paddle board[2] for my birthday.	293	K-4
Critics penned several recent articles.	4	K-2	Critics panned several recent articles.	2	n/a
I just wrecked the pool table.	29	K-3	I just racked the pool table.	1	K-3

Note. Table shows raw frequencies from Spoken British National Corpus (S-BNC) and British National Corpus (BNC) grade level.

[1] Grade level was generated by the online platform, Compleat Lexical Tutor (Cobb, n.d.) based on the complete BNC corpus (100 million words) rather than the S-BNC (10 million words).

[2] Pedalboard and paddleboard tend to be written as single words, but have been written here as two words as the individual morphemes are likely to have been heard. Either option, however, leads to the same frequency band. As pedalboard-paddleboard, the words share frequencies of 0 S-BNC, n/a BNC level.

Word and sentence list development. Word and sentence lists (hereafter, cumulatively "word lists") were read aloud by talkers to create the listening prompts for Experiment 1. Word lists consisted of real rather than nonce words for theoretical and practical purposes. There is evidence to suggest that real words may be susceptible to potential response bias through familiarity effects (Inceoglu, 2022). Consequently, nonce words have been proposed as listeners are theoretically equally unfamiliar with each word. However, nonce words are restricted to English phonotactic rules, and consequently would also be subject to potential frequency effects, both for syllables (Croot et al., 2017) and orthographic representation (Solso & Juel, 1980). Further, because the experiment incorporates mono- and multisyllabic words, the sequence of nonsense syllable order would also be a consideration (Pinker & Birdsong, 1979). In addition to possible hidden frequency effects, nonce words may lead to hyper-articulation (Maxwell et al., 2015), further distancing them from natural speech. Hyper-articulation minimises coarticulation and maximises clarity (Casserly & Pisoni, 2010). When heard amid an otherwise normal sentence, the hyper-articulated word would alert the listener where to focus their attention. Such an alert would have been problematic, particularly for Experiment 2 where participants perceive speech in attention-focused and non-attention focused conditions with Travel Agent and Diverse Sentence prompts. As both nonce and real words carried potential constraints, the option most reflective of the real world (i.e., actual words) was implemented, consistent with the aims of the study.

Experiment 1 word lists were reduced from the pilot based on results, frequency counts, and syllables. Items which did not discriminate between high and low performing participants were excluded. Target word frequency was explored using the Spoken British National Corpus 2014 (S-BNC). The S-BNC was referenced as it was sufficiently large (10.4 million words), reflective of British speech, and publicly accessible. Table displays the word frequency counts from the S-BNC for selected vowels in CVC frames. Target vowels (the first four rows of Table) displayed a minimum frequency of seven (/i, ɪ/ bVd) and a mean frequency range between contrasts of 7009 (*SD* = 10963).

Table 9. Canonical CVC word frequencies

Vowel	bVt	Freq	Grade level	bVd	Freq	BNC level	hVd	Freq	Grade level
/i/	beat	425	K-1	bead	7	K-5	heed	9	K-6
/ɪ/	bit	12317	K-1	bid	418	K-2	hid	36	K-2
/ɛ/	bet	1143	K-1	bed	1839	K-1	head	1905	K-1
/æ/	bat	69	K-3	bad	3076	K-1	had	29316	K-1
/u/	boot	236	K-2	booed	9	K-3	who'd	1141	n/a
/ʊ/	n/a			n/a			hood	56	K-3
/ʌ/	but	6590	K-1	bud	30	K-4	n/a		
/ɒ/	bot	6	n/a	bod	4	n/a	hawed[20] (hod)	2	K-18
/eɪ/	bait	14	K-7	bade	1	n/a	hade	1	n/a
/aɪ/	bite	129	K-2	bide	1	K-11	hide	149	K-2
/əʊ/	boat	627	K-1	bode	1	K-12	hoed	0	K-7
/aʊ/	bout	9	K-5	bowed	3	K-3	how'd	42	n/a
/ɔɪ/	n/a	n/a	n/a	Boyd	5	n/a	n/a	n/a	n/a
/ɜ/	Bert	43	n/a	bird	248	K-2	heard	2385	K-1
/ɑ/	Bart	11	n/a	bard	3	K-16	hard	1957	K-1

Note: Table shows tokens present in Spoken British National Corpus (S-BNC) and British National Corpus (BNC) grade level. Freq = frequency of occurrence in the corpus.

While strict adherence to frequency would be prohibitively impractical for developing listening prompts (evidenced by Table), general frequency categories were created to track potential frequency effects using the number of tokens present in the S-BNC. Categories were high (> 100 tokens), mid (10-100 tokens) and low (< 10 tokens). Applying these categories to target vowels in the canonical CVC frames, only three of six pairs shared the same frequency band: beat-bit, bed-bad, and head-had. All were in the Experiment's "high" category, but with great divergence in individual frequency count given by the S-BNC. Incorporating additional vowels (latter 11 rows in Table) would have exacerbated variance.

Frequencies in the word lists were more closely matched than those found in the CVC frames to ensure that frequency effects would be no greater in the word lists of diverse words than in traditional CVC frames. For Diverse Words, the average difference in frequency of occurrence was

[20] Hawed and hod is displayed for contexts (dialects) where the cot–caught vowel merger has occurred.

615. All word pairs assumed the same band[21], meaning they were approximately equally frequent or infrequent in the Spoken S-BNC. For the Diverse Sentences list, the average difference between pairs was 388. Four of 14 pairs had mismatched frequency bands: three mid-low, one mid-high. Frequency was not used as inclusion criteria for the Directions word list as it employed proper nouns.

Pilot 1 results suggested that the number of syllables may interact with perception; subsequently, an equal number mono- and disyllabic words were included for comparison in the diverse isolated words list. Where possible, words with the same root (provided they were of the same frequency category) were chosen. 'Leave' and 'live', for instance, had disyllabic matches in 'lever' and 'liver', while 'feel' and 'fill' were matches with 'feeling' and 'filling'. No direct comparisons were possible for the street names; however, each vowel pair included a mono- and disyllabic street name (e.g., Sheep-Ship Lane and Siemens-Simmons Road for /i, ɪ/; Fenn-Fann Street and Grenville-Granville Street for /ɛ, æ/).

As sentential context may help in cuing the keyword for participants (Kuperberg & Jaeger, 2016), sentence pairs were created to ensure that semantic meaning would not aid guessing (e.g., 'Let's meet at Redcliffe-Radcliffe Square'). Sentences were generated from personal experience (e.g., 'It will be an expansive-expensive study'), collocations in corpus searches (e.g., 'chromosome banding'), search engine results (e.g., 'Grenville-Granville Street'), or a combination of these (e.g., 'chromosome banding' resulted in a successful search for 'chromosome bending').

[21] minus street names, which did not use frequency as inclusion criteria; see Directions

Recordings. Recordings for Experiment 1 were the same as described in Pilot Recordings.

Target vowel duration. Vowel length was measured to identify characteristics of perceptual targets and to uncover the potential impact of vowel length on perception. Whereas English L1 speakers primarily rely on spectral cues for vowel perception, Mandarin and Korean L1 have been shown to use vowel duration as a primary cue for discriminating between /i, ɪ/ and /ɛ, æ/ (Flege et al., 1997; Wang & Munro, 2009). This was of specific interest because, in addition to typical length associated with each vowel in CVC frames, added linguistic diversity has corresponding variation in duration. Vowel length is inversely correlated with the number of syllables in a word (Ladefoged 2012), so a vowel in a single syllable word will tend to be longer than the same vowel in a multisyllabic word. Further, vowels are shorter when preceding voiceless consonants compared to voiced, when the syllable is closed versus open and when the vowel is unstressed opposed to stressed (all target vowels in the present study carried primary stress).

Measuring duration necessitated marking target vowels in each utterance with subjective boundaries. In naturally spoken speech, there is an absence of boundaries between speech sounds as each segment blends into the next. Not only is there a continuous acoustic signal in fluent speech which physically binds words together, but each speech sound corresponds with a place of articulation. In anticipation of producing an ensuing consonant or vowel, articulators (i.e., the tongue and lips) pre-emptively modify their shape to facilitate the transition from one segment to the next, resulting in neighbouring segments "colouring" each other (see Ladefoged & Johnson, 2014 for an extensive review).

To identify vowel boundaries, visual and acoustic cues were iteratively referred to using the speech analysis software, Praat (Boersma, 2001). Praat provided a simultaneous view of the waveform (oscillogram) and frequencies (spectrogram) of the audio files with text tiers for marking word and phoneme boundaries. These tiers permit users of Praat to make notes on separate layers underneath an image of the audio waveform. Spectral changes in frequency and amplitude identified the target vowels and audio was played to confirm selections. Boundaries were placed at zero crossings[22] where the onset and offset of periodicity[23] occurred. Referencing visual cues for the final measure of duration is congruent with Baart (2010) as the objectivity of visual criteria made it more desirable than the subjectivity of individual perceptions. If a boundary was visually and auditorily unclear, such as when a vowel transitioned to or from a liquid (e.g., the transition between /ɛ/ and /l/ in 'Ellen'), a midway point was selected. Table 1 displays durational differences between vowel pairs.

Table 1. Mean durations (milliseconds) of target vowels in syllables for each word list

		bVt	Diverse Words		Directions[1]		Diverse Sentences	
Vowel	Talker	Mono	Mono	Di	Mono	Di	Mono	Multi
/i/	1	147	221	131	96	100	161	105
	2	167	214	122	82	94	158	114
	3	161	207	153	102	112	170	133
	4	182	219	138	87	86	170	106
	M	164	215	136	92	98	165	114
	SD	12	5	11	8	10	5	11
/ɪ/	1	97	140	59	61	57	103	66
	2	102	121	66	48	42	99	47
	3	137	127	72	48	41	104	57
	4	138	118	78	49	52	98	51
	M	119	127	69	51	48	101	55
	SD	19	8	7	6	7	3	7
/ɛ/	1	154	154	86	105	81	108	81
	2	156	138	70	73	78	115	74
	3	179	152	74	89	75	118	76
	4	196	157	79	81	77	141	80
	M	171	150	77	87	78	120	78
	SD	17	7	6	12	2	12	3
/æ/	1	140	201	106	168	98	188	91
	2	164	200	91	125	94	168	86
	3	211	213	123	106	100	156	99
	4	219	227	95	135	95	177	106
	M	183	210	103	133	96	172	96
	SD	33	11	12	22	2	12	8

[22] Where the waveform is at 0.

[23] Periodicity refers to the regular waveform intervals of the vowel.

Note. ^{1}/i, ɪ/ Directions Sentences consisted of two word pairs: Ship-Sheep Lane and Siemens-Simmons Road. Syllable comparison is confounded because vowels preceding /m/ are longer than vowels preceding /p/.

Table 1 displays an inverse relationship between the number of syllables in a word and the duration of the target vowel (Pearson's r = -.507, p < .001), which is consistent with the literature on vowel compression (Katz, 2012; Klatt, 1975). Because there is a relationship between vowel duration and number of syllables in a word, vigilance should be employed to interpret their respective roles during analysis. Similarly, attention must be given when interpreting single syllable words with frequency, as monosyllabic words tend to be more frequent**.**

Listening Task Design. Listening tasks consisted of six blocks of stimuli: two discrimination tasks and four identification. Tasks are explained immediately below. Each block began with a splash page displaying the block title and instructions, followed by a four-item practice section with immediate feedback (a checkmark for correct, a red X for incorrect) and the formal experiment which did not include feedback. Trials were forced choice in Experiment 1.

Discrimination tasks.

bVt and Diverse Words (isolated words). Blocks 1 (bVt) and 2 (Diverse Words) employed a four-interval (tetrad) oddity design as described in the Pilot (see Pilot for a detailed description and explanation). Results from the pilot study supported the use of four intervals for the diverse prompt types, and in post-experiment questioning, participants reported the number of intervals was not problematic, but knowing the difference between sounds was.

Control trials were randomly included to identify participants who had trouble with the task, whether due to memory or task design. If participants were to systematically incorrectly respond to control items, it would be uncertain whether incorrect responses in the experimental items were due to inaccurate perception or the inability to follow the task. Conversely, if participants correctly responded to control items, but incorrectly to experimental items, it may reasonably be assumed that performance on the experimental items was due to discrimination rather than the inability to perform the task effectively. Control trials contrasted the target vowels with back vowels (/u/ and /a/). Trials with back vowels were expected to be readily contrasted as their vowel spaces did not overlap with the target vowels, and the targeted language groups each contrasted front and back vowels in their L1s.

Instructions were: “You will hear four words spoken by four speakers. Three speakers will say the same word; one will say a different word. Click the button that reflects the odd word. If the odd word is spoken first, click Button 1; if the odd word is spoken second, click Button 2, and so on. Click next to begin your practice”.

Practice for bVt incorporated the words “boot” and “bot”. Diverse Words practice included the pairs “boot-bot”, “coot-cot”, “hoot-hot”, and “pooter-potter”. Each correct answer for the practice items was found in a different position reflecting the roving design of the experiment’s oddity tasks.

Identification tasks. Blocks 3-6 utilised a two-alternative identification task, noted as one of the simplest identification tasks (Winer & Snodgrass, 2015). Participants heard a single stimulus (word or sentence) and indicated which stimulus they heard by pressing one of two buttons. Buttons were each labelled with a single word, reflecting a minimal pair. Using two alternatives (opposed to three or more) is common in perception testing (Klein, 2001) and provided practical benefit for the present study: there are relatively few triads of words which are minimally contrastive, the same part of speech, contextually ambiguous, and similarly frequent. Buttons were dynamically labelled (the words changed each trial) with target words in a non-roving orientation, allowing participants to associate vowels with a specific button and side of the screen. The word *meals,* for example (and all other target words with a stressed /i/ nucleus), was orthographically presented in the left button, while mills (and all other target words with a stressed /ɪ/ nucleus) was presented in the right button. A single stimulus (word or sentence) was aurally presented upon play which participants were permitted to replay once.

Block 3 and 4: bVt and Diverse Words (isolated words). Block 3 employed the canonical bVt frame and consisted of beat, bit, bet, and bat produced by all four talkers and repeated three times, constituting 24 items per pair. Block 4, Diverse Words, consisted of 16 words per pair multiplied by four talkers, totalling 64 items per pair (Gerrits & Schouten, 1998). The instructions were, "You will hear a word or phrase spoken alone (i.e., not in a sentence). Click the button that reflects the word you hear. Click next to begin your practice". Practice tokens were boot and bot for bVt; stop, root, sloth, and potter for Diverse Words.

Block 5: Directions (connected speech). The Directions task was an exploratory design to reflect an applied context for vowel perception assessment. It was a linked speech task that required participants to listen for a London street name. Buttons were screenshots of streets found on Google Maps for Android. Adapting real-world materials was intended to enhance authenticity, lending itself to instructional materials and assessments which promote authenticity. Street names were enlarged in Photoshop 2020 and durations to destination were removed. The play button prompted a recording of, “Meet me at” followed by the destination.

The instructions were, “You are meeting a colleague. Listen and identify the location your colleague wants to meet. Click the appropriate image to get directions”. Practice for this task included the destinations, Bloomfield Road and Blomfield Road.

Block 6: Diverse Sentences (connected speech). As Experiment 1 investigated listening prompt diversity for assessing vowel perception, Diverse Sentences was the ultimate block of interest. It offered the greatest extent of linguistic and phonological diversity within the experiment, flanking the diversity spectrum (with isolated bVt words on the other end).

The instructions were, “You overhear someone talking on the phone. Listen and identify what the person says. Click the button that reflects what you hear”.

Item count. A balanced number of items per block would have been desirable; however, two competing factors prevented such balance. While word pairs were sought which held theoretical potential for conflation and which were roughly equivalent in frequency, this was compromised by the desire for more diversity in syllables and phonological environments. Resultantly, blocks contained disparate numbers of items (to be addressed in analysis). The block bVt identification, for instance, was limited to the four target vowels spoken by the four Talkers, totalling eight unique items for each of /i, ɪ/ and /ɛ, æ/. For Directions, because each street was a local London street opposed to streets in different cities in the UK, there were few minimal pairs available for the high vowel pair. With only four streets matching the inclusion criteria for /i, ɪ/, there were 16 total items, compared to 14 minimally paired streets (totalling 56 items) for /ɛ, æ/. For Diverse Sentences, the sentence, “She’s a little fleshy-flashy” was excluded to avoid the potential for participant provocation or offense, resulting in eight items being removed for /ɛ, æ/. No replacement pairs of (approximately) equal frequency were available. The full breakdown of items is in Table 2.

Table 2. Item count by type and vowel pair

Task Block	/i, ɪ/ items	/ɛ, æ/ items
bVt Oddity	32	32
Diverse Words Oddity	64	64
bVt Identification	8 (x3)	8(x3)
Diverse Words Identification	64	64
Directions	16	56
Diverse Sentences	56 (8 excluded)	64

Open-ended question

The experiment included a qualitative element in the form of a voluntary, open-ended prompt at the end of the data collection session. The textbox prompt stated, “We'd like to hear how you felt about the different parts of this experiment, and welcome comments below. To refer to individual parts of the experiment, here's what you did (not necessarily in order)”. Each section was labelled with a number and descriptive title, followed by an example.

Though a single item can be expected to yield an impoverished response compared to a formal interview, I reasoned that participants who undertook the item would include most pertinent aspects of their experiences, given by their relative saliency (Geer, 1991). The prompt provided me with an expedient means to obtain exploratory, useful data while respecting participants' uncompensated time. In Experiment 1, 28 participant responses formed a corpus of 1505 words.

Following Braun and Clarke (2006), I applied a thematic analysis to participant responses using an iterative approach, which included familiarisation with the data; generating and making inferences; grouping codes into themes, developing a preliminary thematic map; ensuring coherence between themes, the original data and the underlying theoretical perspective; formalising a final thematic map; defining the themes and their relevance; and reporting the findings. (For a comprehensive account of these phases, see Terry et al., 2017.) To avoid a procrustean template for viewing the data, I used an inductive (bottom-up) approach. Such a data driven approach is suitable for developing coding schemes which are theoretically more congruent with the original content of the responses (Braun & Clarke, 2012; Terry et al., 2017[24]).

For enhanced reliability, I conducted a pilot coding with an external coder, then formal coding with a second external coder. A focus on reliability supported my ontological (post-positivist) perspective of the data and its analysis, but was contra to Braun and Clarke's perspective. I justified utilising their analytical framework because it provided a logical, systematic roadmap for analysis; had been incorporated in recent years by leading researchers in relevant fields, such as speech perception (Harding, 2017), psychology (Clarke & Braun, 2018), and education (Xu & Zammit, 2020); and was not intrinsically associated with any one theory (e.g., grounded theory) (Terry et al., 2017).

[24] Tyler et al. explain that detailed engagement engendered within an inductive approach promotes effective coding through immersion (p. 6).

Code development. Familiarising myself with the data, I read the text responses several times and designed a working code structure on paper, complete with overarching themes before transferring codes to the software programme, NVivo 12 (QSR International, 2018). Initial coding was iterative as I applied the codes to the data, refined (clustering and collapsing themes and codes), and applied again. After making notes and definitions to specify what the codes were and how to use them, I recruited a PhD student with professional experience in applied linguistics and language testing as the pilot coder. I then trained the coder in an hour-long synchronous video chat session where I explained code definitions and we practiced applying codes on selected excerpts. Excerpts were chosen extemporaneously to demonstrate code use and coverage. Upon completion, the pilot coder independently coded the 28 responses.

Pilot coding results suggested I needed to refine codes and training before the formal coding. The initial code structure offered a comprehensive, inductive reflection of the original data, but included superfluous information (e.g., experimental design, study utility) which did not address the study purpose of identifying prompt specific information to help identify strengths and weaknesses of traditional and diverse prompt types; data reflecting the overall experimental design or perceived utility were external considerations, but present in the pilot code structure. This additional information unnecessarily complicated the coding structure, which in part led to another problem: unused or sparsely used codes.

Reviewing the coding and reflecting on the training process, I identified areas for improvement in both coding and training. The codes were too numerous and occasionally redundant (e.g., “easy” and “difficult”, opposed to “difficulty”), presenting a tree of codes that was challenging to reliably implement. Insufficient training exacerbated this problem. There were too many codes to make it practical to explain and use each in training, and more time was needed to learn each code, the differences between codes, how to apply them effectively, and how much text to include.

Stemming from the pilot results, I created a formal, simplified coding scheme with fewer codes (e.g., excluding codes related to experimental design and study utility) and more direct focus on the research question. I also refined the research question from focusing on general participant experience with each prompt type to participant cognition with each prompt type. “Cognition” consisted of any feedback which implicated how participants thought or felt regarding specific prompt types. With these revisions completed, I developed a PowerPoint presentation for more thorough, systematic training which would yield more robust results (Appendix: Experiment 1 Coder Training).

With the revised coding, I printed the data and cut participant responses into separate strips. This permitted a more tactile, visual means of coding the data and helped me concretise the coding process. Using a pen, I wrote codes on each strip and then sorted the strips into categories (e.g., Attention). Where a strip was coded in two categories or themes, I placed it between categories. Satisfied with the coding scheme, I recruited and trained a second external coder (Coder 2) for the formal coding.

Coder training (and further code refinement). Coder 2 was an applied linguistics researcher and assistant professor with experience coding related data. Training consisted of a pre-recorded PowerPoint presentation, phone conversations, and calibration exercises. The presentation included background information about the study, requirements for the coding, vocabulary specific to the study (prompt types and tasks) that participants commonly referenced, codes, and a coding review quiz to familiarise the coder with each code in context. I was on hand to answer questions during and after the training presentation. Upon completion of the training, Coder 2 applied his training on an NVivo calibration file which included real data from 4 excluded participants.

I separately completed the calibration data using the revised codes, and after Coder 2 submitted the coded calibration data, I manually compared our two codings. I went through each code, comparing similarities and differences, and documenting them in a Microsoft Word file. I highlighted key areas of disagreement and error, and added explanatory notes to reconcile discrepancies. During this process, I found opportunities for improving code accuracy. Some cognitions were present in the data, but not in sub-codes taught to the second coder. For example, references to participants' ability to focus could reflect Attention, a parent theme in the code structure, but not sub-codes which affected Attention (e.g., Fatigue). Parent themes, then, became parent codes which we could use for general reference to the theme (i.e., beyond specified sub-codes). Similarly, to account for the possibility that Coder 2 may find participant cognitions which were unaccounted for in my coding scheme, I added a new code, "General Cognition". This last code was a "wild card" and, with its broad scope, threatened reliability; however, this threat was counterbalanced by providing a means of flagging data I had not originally considered. The new, more general codes offered a built-in placeholder for asynchronous discussion.

I sent the completed reference document to the second coder, and we again spoke over the phone to discuss differences. Once we were satisfied with our understanding of the codes and how they were to be used, Coder 2 continued to independently code the real, full dataset and I re-coded based on the revised coding system. Once the second coder had double coded 100% of the data set, I reviewed the coding (along with my own) and excluded all codes which did not directly indicate a prompt. Removing non-prompt-specific coding meant fewer codes specifying cognition, necessitating a modest code structure revision. Memory, which was left with zero references, was omitted, while other codes were regrouped. With coding already completed, it was not suitable to change the coding structure for calculating reliability; however, codes which had too few individual references were assimilated into others after reliability analysis was completed. Specifically, Fatigue and Emotion (Positive and Negative Affect) were grouped with General Cognition. I then calculated inter-coder reliability, as described in the section, Methods Reliability.

Table 3. Experiment 1 codes, definitions, and examples used by coder

Code	Code definition	Data example (participants' verbatim comments)
Attention\focus and general attention	Reference to focus or attention which is not encompassed by fatigue, memory, or confusion.	*not very challenging but take effort when I am not 100% attentive*

Attention\fatigue	Exhaustion, tiredness or weariness caused by specific prompts	*when I hear these words reappearing over and over again, I felt less concentrated and I misclicked one of the choices*
Confusion	Uncertainty, particularly that which may misdirect or effect a participant's attention.	*some sounds got more confused for me, since i do not even recognize their differences when i use them in my daily conversations*
General Cognition	Text which is relevant to cognition, but has no other code to describe it. Key examples include familiarity or miscellaneous thoughts about a prompt (or how the participant approached it).	*when put into the sentences, some sounds are more easier to be recognized like meals and mills, maybe it is because i have some more time to prepare.*
Perceived difficulty	Any reference to difficulty, easy or difficult. This is perceived because participants may perceive a task as difficult but perform well, or perceive a task as easy but perform poorly.	*Choosing the odd word out (discrimination tasks) felt easier than the later parts of the experiment (e.g., selecting buttons labelled with street names).*
Strategies	Methods participants use to answer the prompts beyond listening perception. Examples include using context to answer a question or guessing.	*i sometimes tend to choose the word that i feel the right in the sentences inevitably rather than fully concentrating on what i've heard*
Prompt\bVt oddity	Reference to the bVt Oddity prompt type	*b-vowel-t Discrimination is a bit hard. bet and bat, I can't really tell.*
Prompt\Diverse Words	Reference to Diverse Words prompt type	*The bvt-Identification is easier than the Diverse Words-Discrimination for me.*
Prompt\Directions	Reference to Directions prompt type	*Diverse Words-Discrimination: It felt harder to distinguish words that I wasn't familiar with*
Prompt\Diverse Sentences	Reference to Diverse Sentences prompt type	*Diverse Sentences: Overall okay, but there are few times where the pronunciation sounds very much like both the options given*

Reliability. As the coding application was a subjective process, consistency (reliability) of the coding was a relevant concern (Haertel, 2006). Intercoder reliability—the extent to which coders agree on the application (and non-application) of codes for a qualitative dataset—was examined to promote systematicity, transparency, and rigor through dialogue (O'Connor & Joffe, 2020). The study employed two quantitative measures of intercoder reliability: absolute percent agreement and Cohen's kappa. Together, percent correct and kappa offered a more complete, readily interpretable indication of agreement than would be possible independently. Absolute percent agreement provided a descriptive index of agreement, measured by the number of times coders agree, divided by the total number of ratings. Percent agreement is practical in its simplicity, but is vulnerable to chance inflation when coders are uncertain (Cohen, 1960). I thus applied an additional measure, Cohen's kappa (k), to correct for chance. Kappa incorporates chance with percent agreement, as shown by its formula:

$$K = \frac{Po-Pc}{1-Pc} \qquad (1)$$

where Po is the proportion of agreement between coders and Pc is the probability of chance agreement. Though kappa is a more sophisticated measure of agreement, it is also less intuitive to directly interpret than percent agreement. Calculating kappa produces an output between -1 and 1. Values between -1 and 0 indicate disagreement; positive values indicate agreement, with 1 demonstrating perfect agreement. There is no universally accepted interpretation of specific kappa values, but Landis and Koch (1977) offer a guideline for ranges, where 0.21-0.40 reflects fair agreement, 0.41-0.60 is moderate, 0.61-0.80 is substantial (the target minimum overall range for the present study), and 0.81-1.00 approaches perfection. Anything less than 0.21 is slight or poor.

Percent correct and kappa both calculate intercoder reliability by aggregating results for each code; however, aggregated results can obfuscate individual discrepancies in ratings, particularly where disagreements arise. Therefore, displaying individual computations for each code was appropriate for transparency and to help interpret findings.

Procedure

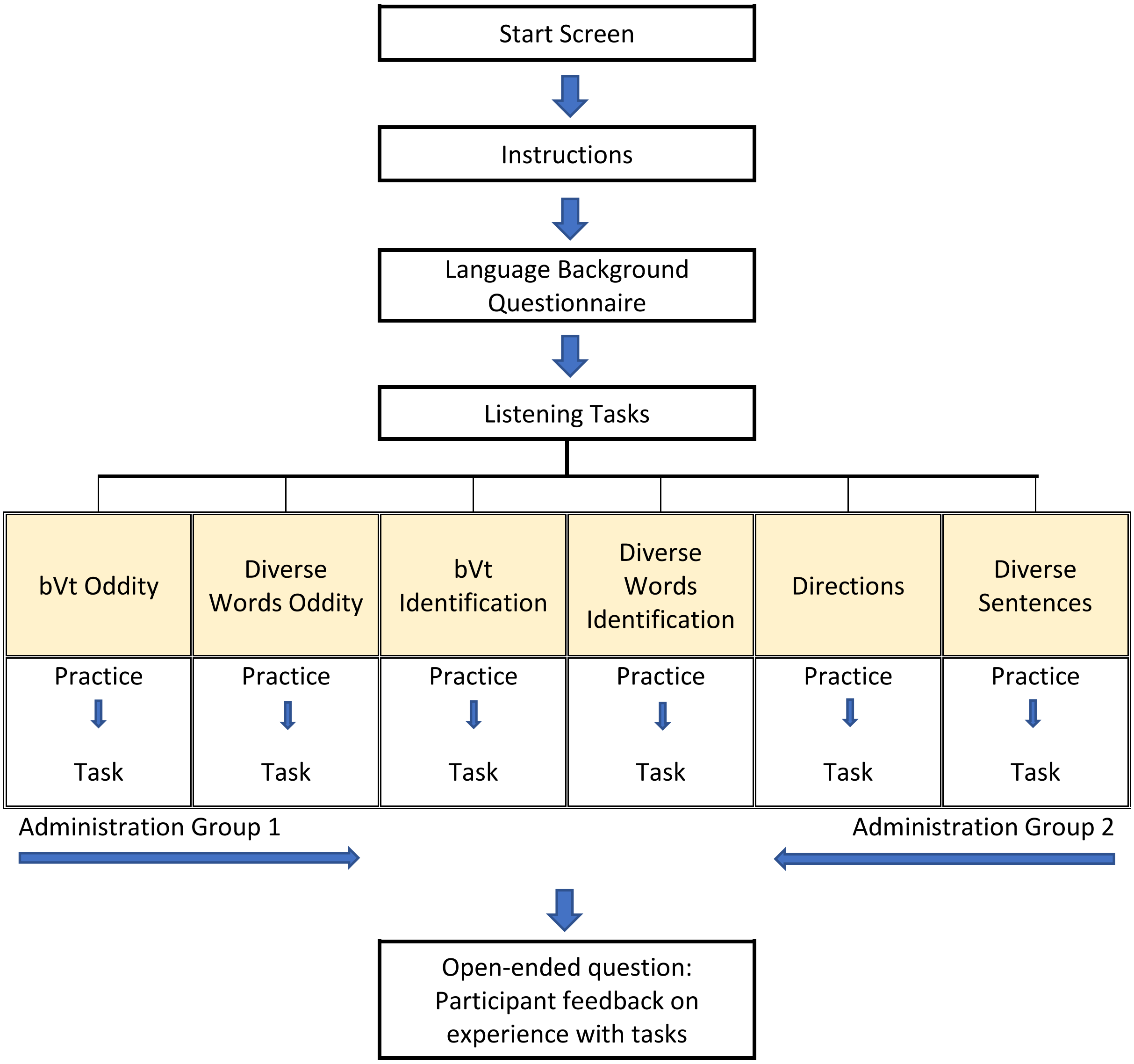
The study was conducted in a quiet room at the university in a single session (approximately 90 minutes, including breaks). The experiment was administered using Gorilla.sc (Anwyl-Irvine et al., 2020), an online platform for building and administering behavioural experiments. Immediately prior to the listening experiment, I explained the study's tasks and procedures, and that I would be present throughout the experiment. Participants completed a paper-based consent form and language background questionnaire, and were then given a pair of padded on-ear headphones. Before following the link to the experiment, participants were asked to listen to the start of a Youtube music video at their assigned computer to confirm their headphones were properly connected (a problem encountered in the pilot) and to adjust their headphone volume as needed. Participants then clicked on the official experiment link and began the experiment (see **Figure 1**).

There were six listening tasks where stimuli were presented binaurally and in random order. Tasks were self-paced to enable audio replay (see Replay). An automatic, fixed interval advancement would have ensured presentation uniformity, but the platform would not permit replay and fixed interval timing to co-occur. A progress bar allowed participants to monitor their progress within that particular task and a message notified participants of the midpoint of each task.

The experiment was administered to participants in one of two sequences, creating two groups. Group 1 received the blocks in order bVt Discrimination-Diverse Words Discrimination-bVt Identification-Diverse Words Identification-Diverse Sentences-Controlled Sentences; blocks were counterbalanced for Group 2 with participants completing blocks in the opposite order to Group 1. The pilot showed no sequence effects, and none were expected for this experiment.

After all tasks and their associated surveys had been completed, participants were given the opportunity to write additional notes about their experience in a textbox. For participant reference, the question listed each of the item types with examples.

Figure 1. Graphical outline of design and procedure for Experiment 1



Note. Tasks were administered in opposite order depending on group. Participants were randomly assigned to group administrations.

Analysis and statistical approach

RQ1. Compared with bVt prompts, to what extent are more phonologically diverse prompts (diverse isolated words, words in a fixed carrier sentence, words in syntactically diverse sentences) suitable for assessing English vowel perception in advanced L2 learners at a London university?

Research questions were subdivided to facilitate systematic exploration. Each are described subsequently with their respective analyses.

Q1a. Compared with bVt, prompt to what extent are diverse prompts reliable measures? This was explored with Cronbach's alpha, a measure of internal consistency. In L2 speech perception studies, reliability is routinely unreported. A search of peer reviewed HPVT studies published between 2018-2021 and accessible by the databases, ERIC, ProQuest, Scopus, and Web of Science, yielded no entries of internal consistency of items. A single reporting of reliability was found for inter-rater reliability[25]. Reliability was deemed necessary for the current study because (a) it is consistent with best practices amidst the larger spectrum of language testing, (b) it promotes transparency and accountability by reporting internal consistency against an a priori criterion, and (c) it permits a useful statistical comparison of tasks.

Cronbach's alpha of 0.7 and above is often considered acceptable, with lower levels justified for scales with few items (Taber, 2017). There are varying interpretations, however, and a range of 0.4-0.55 may be considered as either acceptable or not satisfactory (see Taber, 2017, for a comprehensive explication of reliability range judgments). The criterion of "acceptable" for the current study is 0.7 for most tasks and groups, with latitude given to the exploratory Directions Sentences' /i, ɪ/ vowel pair which presented only 16 items.

[25] Dong et al. (2019) published a peer reviewed paper featuring high and low variability training which required coding pre- and post-test results. In addition to Dong, a dissertation from Isbell (2019) was found with the search terms, "hvpt" and "reliability".

Cronbach's alpha is sensitive to item count. Higher numbers inflate alpha, while lower numbers deflate it. Given there were uneven numbers of items across the listening blocks, the number of items for each task was reduced to 32, the lowest common item number across prompt types which could be measured with Cronbach's alpha. BVt Identification prompts were excluded from analysis due to item repetition (items were repeated to bolster the block of items as there were only four bVt [i.e., beat, bit, bet, bat] spoken by the four talkers). Items for were randomly chosen for exclusion using the RANDBETWEEN function in Excel. Numerals 1 and 0 were randomly assigned to each item, with items marked 1 selected for exclusion. A random selection was chosen over a targeted selection of the poorest performing items (e.g., extremely difficult or easy items; items which discriminate between high and low performing participants) to avoid artificially inflating the statistics from cherry picking the data[26].

Q1b. Compared with bVt, to what extent do diverse prompt types match predictions of group performance for Mandarin, Korean, and control? To test this, first the control group was used as a baseline to identify performance levels of individuals predicted to readily discriminate between target vowels (Ingram & Park, 1997). These results were compared with each group's expected performance. Within-group performance was measured with percent correct and a dependent *t*-test investigated the significance of the difference between vowels pairs. Between-group performance was explored in mixed models design (see Q1d. To what extent do bVt prompts predict performance with Diverse Sentences? for model design explanation).

A mixed models design was selected over analysis of variance (ANOVA) to compare language groups due to its ability to account for non-independence, differing levels of item difficulty, predicted person ability, and disparate levels of variance among groups. As the control was expected to perform at or near ceiling, comparatively little variance between participants was anticipated, while the Mandarin and Korean L1s were expected to display a range of scores—some performing

[26] Omitting poorly performing items would be desirable for developing a customised dataset for training and testing specific populations in a real-world setting; however, it would be misleading to have it here, a measure of what you might generally (opposed to ideally) find with a given number of items.

well and others at chance levels. Unequal variances and unequal sample sizes can result in increased Type I error rates (Rusticus & Lovato, 2014). Unlike ANOVA, mixed models are more robust to such factors (Bolker et al., 2009). Details of the mixed model design are elaborated in Q1d. To what extent do bVt prompts predict performance with Diverse Sentences?.

Performance predictions and PAM-L2. Establishing how individual language groups perform with the fixed frame (bVt) prompts permitted inferences regarding the effects of prompt diversity and outlined limitations of the data. Predictions were made for both within and between group performances based on PAM-L2 (see Literature review) and empirical evidence.

Mandarin performance predictions were motivated by the Mandarin phonemic inventory along with discrimination and identification studies. The precise vowels which make up the Mandarin vowel inventory is debated (Sun & Van Heuven, 2007), complicating initial predictions. Lee and Xiong (2021) explain that the difference arises for two reasons: first, because of variations in the high mid, unrounded vowel (i.e., /ɨ/); the second, and of pertinence to the current study, due to the front low vowel. According to Lee and Xiong, "The low vowel also has three allophonic variants [a], [ɑ], and [ɛ]. The phonemicization of the mid and low vowels largely varies in the literature" (pp. 332-333). To Lee and Xiong, Mandarin front vowels consist of /i, e, a**,** y, ɨ[27], u/; as the present study is strictly concerned with front (unrounded) vowels, the first three (/i, e, a/) are the focus. A different set of front vowels is proposed by Huang and Liao (1997), with /i, ɛ, a/—where /ɛ/ replaces /e/—and Li (1999) includes both front mid vowels with /i, e, ɛ, a/. The exact constitution of the vowel system can lead to different predictions and reference to the literature on Mandarin L1 acquisition of English vowels was required for refinement.

Jia et al. (2006) compared Mandarin monolinguals, recent arrivals to the US, and "past" arrivals' discrimination of English vowel pairs with an AXB[28] task (along with a production component

[27] Lee and Xiong wrote this as [ɿ, ʅ] (p. 332), reflecting a high, central, unrounded vowel. It is Pinying rather than IPA. The corresponding IPA vowel is ɨ, as shown.
[28] In AXB designs, listeners hear three stimuli. The first and third reflect the contrasting stimuli in a minimal pair, and listeners must decide whether the second stimulus is more like the preceding or succeeding stimulus.

which is outside the scope of this study). Framing their study within the PAM framework, the researchers posited that the /ɛ, æ/ vowels' close proximity in vowel space and lack of phonological counterparts in Mandarin would translate to Mandarin L1s assimilating the vowel pair as SC (p. 1121). The noted lack of counterparts suggests that the researchers assumed a front vowel inventory similar to Lee and Xiong (/i, e, a/). Though /i/ and /ɪ/ are close in vowel space, English /i/ has a direct counterpart in Mandarin while /ɪ/ does not. Jia et al. thus surmised that /i, ɪ/ would be assimilated as CG. Since PAM-L2 predicts CG vowel assimilation will yield better results at differentiating contrasts than SC, /i, ɪ/ should result in higher accuracy rates than /ɛ, æ/.

Relative and absolute performance reported by Jia et al. offer different conclusions for the accuracy of PAM predictions. Going by relative performance, where TC > CG > SC, results supported predictions for the front vowel contrasts (see Appendix: Table of results PAM study of Mandarin L1), though it should be noted that performance for both pairs ranged from chance levels to 100% accuracy. Absolute performance, however, was more contentious as participants performed considerably stronger than expected given PAM (and PAM-L2) descriptors. PAM associates SC with 'poor' performance (Tyler et al., 2014, p. 4), and 'fairly poor' described as performance at approximately 70% (p. 9). Further, CG is associated with good (80-90%) to excellent (>90%) discrimination. In Jia et al., the /ɛ, æ/ contrast was SC, yet yielded discrimination scores of 76%, 89%, and 92% for monolinguals, recent arrivals, and late arrivals, respectively. PAM-L2 allows for improvements with experience, yet if SC discriminates poorly and poor equates to 70%, the monolinguals are above what is expected for SC. It may be surmised that either the model descriptors are off or the experiment's design led to an inflation of scores[29].

[29] Two explanations may help account for Jia et al.'s inflation of scores. The researchers employed a single talker for recording stimuli and this may have helped listeners use talker-specific alternations to discriminate between vowels. Because there only a single talker was used, it is conceivable that participants were able to use non-spectral cues to differentiate between stimuli, such as vowel duration. This would have been particularly salient as a single consonantal frame (dVpə) was utilised. The lack of consonantal and syllabic diversity would have conceivably constrained the speaker from producing instances of each vowel with different durations.

Results from Thomson (2012) complement Jia et al.'s discrimination findings. Thomson tested, trained, and tracked Mandarin L1 English vowel identification performance over eight sessions. Prior to training Mandarin speakers performed best with /i/ (accuracy of approximately 75%), but had trouble identifying the vowels /ɛ/, /æ/, and /ɪ/ (each around 20%). By the third session, performance had improved to roughly 40% for the /ɛ, æ/ vowels, and 65% for /ɪ/. Scores improved further with additional sessions; however, these initial sessions illustrate that with some level of intensive exposure (i.e., L2 speakers are no longer naïve to the target vowels), we may expect Mandarin speakers to struggle with /ɛ, æ/, and less so with /i, ɪ/.

Predicting Korean assimilation of English vowels was more complicated than for the Mandarin group. Contemporary Korean has a monophthongal vowel inventory of /i, e (ɛ), a, ɨ, u, o, ʌ/ (Lee & Jongman, 2016). The Korean vowel inventory includes phonemic counterparts to English /i/ and /ɛ/, but not English /ɪ/ and /æ/ (Shin, 2015). Given the close proximity of /ɪ/ with /i/ and /æ/ with /ɛ/, the "new" vowels may be predicted to assimilate to the existing categories as CG, as described by Jia et al. for Mandarin. However, the phonemic inventory of contemporary Korean vowels is influenced by traces of historical features which are no longer contrastive (Sohn, 2015). Affecting CG predictions for the front mid-low vowel pair, /e/ and /ɛ/ have merged in younger generations[30], resulting in a larger range of variation for the /e/-/ɛ/ vowel space (Ingram & Park, 1997, p. 348). As English /ɛ/ and /æ/ have considerable overlap in vowel space (Barrios & Hayes-Harb, 2021), the overlapping may interact with the larger vowel range to yield SC assimilation rather than CG. The empirical evidence of Korean L1 discrimination of /ɛ, æ/ supports the vowels assimilating into a single category in Korean (Flege, 1995a; Ingram & Park, 1997), making it challenging for Koreans to distinguish between English front mid-low vowels.

For the high vowel pair, as /i/ is present in Korean, it offers a target for assimilating English /i/, and /ɪ/ may be predicted to assimilate to the same /i/ category due to proximity (Ingram & Park,

[30] Korean /e/ and /ɛ/ have merged in contemporary Korean (Shin, 2015), yet may still exist in older generations and remain distinct graphemes in Hangul, the Korean writing system.

1997). How good of an exemplar /ɪ/ is determines whether it is SC or CG. If not a good exemplar, it will be CG assimilation leading to stronger predicted accuracy scores for /i, ɪ/ than /ɛ, æ/. If it is a good exemplar of the /i/, assimilation will be SC as both English /i/ and /ɪ/ would be strong representations of Korean /i/.

Support for /i, ɪ/ as CG is found in Ingram and Park (1997) in their investigation of the influence of Korean (and Japanese) L1 phonology for identifying Australian English vowels (/i, ɪ, e[31], æ, a:/). The researchers found that both inexperienced and experienced Korean participants identified /i/ and /ɪ/ more effectively than /e/ and /æ/. Both groups identified /i/ with 100% accuracy. The inexperienced group accurately identified /ɪ/ 82% of the time (mistaking it for /i/ 16% and /e/ 2%) compared to the experienced group's 72% (28% of the time mistaking it for /i/). For the mid-low vowel pair, the inexperienced group identified /e/ correctly in 50% of the trials, but as /æ/ in 48%. They identified /æ/ correctly in 54% of trials and as /e/ in 46%. The experienced group identified /e/ correctly 90% of the time and as /æ/ 10%; they identified /æ/ correctly 76% of the time, and as /e/ 24%. While aligned with predictions, Ingram and Park suggested that improvements in scores between groups may have been confounded by age, as the experienced group was, on average, 6 years older than the inexperienced group and may have had more potential for exposure with an /e, ɛ/ contrast. There is a historical distinction between Korean /e/ and /ɛ/ which remains at least in older speakers, and consequently, there may have been a "residual" L1 influence in the study's results, permitting an additional target for assimilation of /æ/ (p. 354).

Though Ingram and Park's data fit performance expected for CG for /i, ɪ/, and SC for /ɛ, æ/, not all studies support this. Flege (1995b, as cited in Flege 1995a), for instance, found that experienced Koreans did not reliably discriminate between either /i, ɪ/ or /ɛ, æ/. Flege posited that Koreans have trouble with the English /i/ and/ɪ/ contrast because they are associating it with a (now merging) Korean length contrast between /i/-/i:/. As the distinction is no longer strictly contrastive (Kang et al., 2015), it results in a "free variation" allophonic relationship. Flege, explains, "We might

[31] Australian English has /e/, which is viewed here as analogous to /ɛ/ given the Korean merge of /e/ and /ɛ/.

speculate that it is especially difficult for non-natives to discriminate two L2 vowels if phones resembling realizations of the L2 vowels occur in free variation in the L1" (p. 251). An established segmental discrimination researcher reporting similar results for both English vowel pairs complicates predictions which may be made. It is unclear whether /i, ɪ/ should be assimilated as CG, SC, or perhaps a transitional distinction occurring as a result of learning.

Recent work from Lee and Cho (2020) helps clarify the /i, ɪ/ predictions through a series of tasks investigating how Koreans map Standard Southern British English (SSBE) and American English vowels to Korean L1 representations. SSBE, opposed to American English, will be reported here for pertinence and brevity. Participants were upper intermediate to advanced English speakers, split into groups based on short (*M* = 4 years) and long (*M* = 11 years) length of residence in the US. The longer LOR group primarily consisted of participants who had an age of arrival of 12.3 years (*SD* = 5.7), ranging from 4-21 years, and is not comparable to the present experiment's sample. It will therefore not be elaborated. The short LOR group had an AOA of 22.8 years (*SD* = 5.8), and though the range included a participant who arrived in the US as young as 10 (range 10-27 years), it remains the closest analogue to disambiguate predictions for the current study. Participants had limited experience with British English, having reported being taught with American varieties in Korea.

Lee and Cho found that when ascribing a Korean (Hangeul) label to English vowels in an isolated bVt context, participants associated English /i/ and /ɪ/ with Korean /i/ 100% of the time. The investigators asked participants to identify goodness of fit for these vowels on a 7-point Likert scale, with 1 being "very different" and 7 being "very similar". Participants rated the /i/ goodness as 5.1 and /ɪ/ as 5.2. Together, the 100% association of both SSBE vowels with Korean /i/ and the similar goodness ratings suggest SC assimilation, where the non-native vowels are equally good representations of the L1 category. A further identification task revealed these participants readily discerned whether a vowel was /i/ or /ɪ/, achieving accuracy rates of 86% and 75% for /i/ and /ɪ/, respectively. Hence, despite the vowels being equally good representations of Korean /i/, it would seem that for the isolated bVt context, these advanced participants demonstrated the ability to separately categorise these L2 vowels, lending to the notion of TC. There must be an asterisk placed on this label of TC, however, as durational cues between the vowels could well be used by participants in this limited context. Indicated by Lee and Cho, /i/ was approximately 25% longer than /ɪ/ in their study. The advanced proficiency of the participants combined with years of residence in a native English environment may reasonably prepare them for using such a cue. It is therefore unclear whether participants had developed a new category to assimilate the L2 vowel to or whether they simply relied on duration to disambiguate otherwise ambiguous phones. Regardless of how a distinction is made, whether TC or CG, results from Lee and Cho and Ingram and Park suggest that one phone appears to be a better exemplar than the other, leading to more accurate perception than SC. Flege's findings were part of an unpublished manuscript and may have been task or participant dependent; not enough information is provided to make a determination.

Having gathered factors which may influence Korean predictions and explored a study which investigated how well intermediate to advanced speakers identify SSBE vowels, it was expected that Korean participants would show poor to modest performance for the /ɛ, æ/ contrast and better performance with the /i, ɪ/ contrast.

The control was expected to categorise the target vowels pairs distinctly (akin to what TC predicts), leading to a prediction of ceiling performance. This prediction is empirically supported by control results from Flege (1994), Ingram and Park (1997), and Tsukada et al. (2005).

Cumulatively, within-group predictions of group performance were summarised as:

Control (TC) > Mandarin and Korean) /i, ɪ/ (CG) > Mandarin and Korean /ɛ, æ/ (SC).

Additional to within-group performance, the study also attempted to predict between-group performance. In a study engaging (and comparing) all groups in the present research, Flege, Bohn, and Jang (1997) explored the effects of English experience on perception and production of English /i/, /ɪ/, /ɛ/, and /æ/ in a synthetic (computer generated) bVt context. Language groups included Mandarin, Korean, and English L1 (as well as German and Spanish). Participants identified vowels in *beat-bit* and *bet-bat* continua. The experienced Mandarin group outperformed the experienced Korean with /i/ (84% to 60%, respectively) and /æ/ (77% to 43%); the inexperienced Mandarin group outperformed the inexperienced Korean group with /i/ (80% to 75%, respectively), /ɪ/ (83% to 61%), and /æ/ (58% to 18%). (in addition to the experienced Korean group for /i/ and /æ/ vowels). The English L1 obtained perfect scores for /i/, /ɪ/, and /æ/, and achieved 99% identification accuracy with /ɛ/.

Beyond the vowel identification study by Flege, Bohn and Jang, there is a paucity of research which directly compares Mandarin, Korean, and English L1 perception of English vowels; however, in an unpublished study serving as a precursor to this doctoral research, Jones (2015) investigated Chinese (Mandarin and Cantonese), Korean, Japanese and English L1 discrimination of target vowels in /hVd/ and diverse word contexts (mono- and multisyllabic real words other than hVd and nonce words). Preliminary results showed a significant effect for language group, *F*(2, 56) = 5.2, *p* = .08. η^2 = .16. Scheffe's post hoc analysis displayed a statistically significant difference between Japanese (*M* = 212.33, *SE* = 8.23) and the other two target groups, Korean (*M* = 178.44, SE = 6.71) and Chinese (*M* = 188.84, *SE* = 3.04). The mean difference between Korean and Chinese groups was non-significant (*p* = .38). Added to Flege et al. (1997), in side-by-side comparisons there is a trend that Mandarin tends to display higher accuracy scores for the target vowels than Korean, though the significance and magnitude of the differences remain negligible.

Q1c. To what extent does prompt-level complexity affect listener performance? Here, prompt-level complexity refers to the phonological and sentential diversity between prompt types. The bVt prompts (isolated, monosyllabic CVC frames) contained the least variability among all prompt types. This was followed by isolated mono- and multisyllabic real words, then location names at the end of a fixed carrier sentence (Directions), and finally sentences of varying lengths, syllables, syntax, and words (Diverse Sentences). The effect of prompt-level complexity was investigated by comparing participant performance with each prompt type. Percent correct and *d* prime (*d*') were selected to represent common measures of perceptual accuracy in speech perception studies (Nagle & Baese-Berk, 2021) [32]. Whereas the previous research question investigated performance by language group, this prompt-level complexity question collapsed the L2 groups (at the exclusion of the control) to explore overall performance by prompt.

Percent correct (also presented as proportion correct) is a widely used index of performance for vowel perception studies (e.g., Barrios & Hayes-Harb, 2021; Flege, 1995a; Iverson & Evans, 2009) and permitted a descriptive, readily understood means for comparing groups and prompt types. The statistical significance of the differences was probed using a repeated measures ANOVA. Percent correct was reported alongside *d*' (Hazan & Simpson, 2000) to facilitate interpretation.

D', used to account for response bias, is a more sophisticated measure than percent correct and requires elaboration. Originally derived from signal detection theory for identifying how well the presence or absence of a signal has been detected, the measure is readily converted to speech identification tasks with two options (McGuire, 2010), making it ideal for the study's identification tasks[33]. Sensitivity relates to the strength of the signal and its interaction with bias or strategies. A

[32] Naegle and Baese-Beck (2021) outlined common measures of accuracy, but did not describe *A* prime (*A*'). *A*' has been used as a non-parametric counterpart to *d*' and would have matched the binary correct-incorrect responses of the experiment. However, *A*' has been associated with confounding ability and bias, with Pastore et al. claiming that *A*' over *d*' justifications are based on "distorted caricatures" of signal detection theory (Pastore et al., 2003). Given push-back against *A*' and the historical and contemporary prevalence of *d*' for analysing data similar to that in the current study, *d*' was selected as the standard.

[33] Oddity tasks were excluded from comparison as it was not designed to include catch (change no-change) trials (necessary for calculating false positives).

strong signal (or demonstrated ability to detect it) reflects an easy task, a weak signal a difficult task. Figure 2 shows the response matrix used for calculating *d’*.

Figure 2. Response matrix for signal detection

	Response	
Signal	“Yes”	“No”
Present	Hit	Miss
Absent	False alarm	Correct rejection

The number of “hits” and “false alarms” creates two response distributions[34]; the difference between those distributions is the *d*’ index. *D*’ is obtained by subtracting the *z*-score of the false alarm rate (proportion of false alarms) from the *z*-score of the hit rate (proportion of correct hits). The larger the difference between the two distributions (i.e., the less overlap), the more a participant shows the ability to detect differences between target sounds. *D*’ indices at or approaching zero indicate chance scores, while 3.0 suggests near perfect detection. Macmillan and Creelman (2005) suggest the desirable range of scores is between 0.5-2.5. It is uncommon for participants to obtain perfect scores (no overlap between hit and false alarm distributions) as signal detection typically investigates decision making in conditions of uncertainty. Because *d*’ uses z-scores (a standard score), it enables direct comparisons across measures (i.e., prompt types). It was expected that increased complexity in prompt type would have a corresponding decrease in *d*’.

Q1d. To what extent do bVt prompts predict performance with Diverse Sentences?

Isolated bVt prompts (identification and discrimination) have been classically used to determine how well participants identify and discriminate between phonetic contrasts. How well this may translate to sentential contexts is unclear. It was therefore of interest to explore the extent to which bVt identification and discrimination prompts may predict performance on Diverse Sentence prompts.

[34] There are four total distributions (hit, correct rejection, miss, false alarm), but d’ is the difference between hit and false alarm distributions. See MacMillan and Creelman (2005) Chapter 9 for detailed explanation.

Predictivity of bVt prompts was compared with other prompt types in a series of generalised linear mixed models (GLMM). The GLMMs were constructed with the lme4 package (Bates et al., 2015) for the R programme (R Core Team, 2021). Target vowel pairs were investigated separately. Obtaining a correct response was the dependent (outcome) variable. As the outcome variable was binary, a binomial link function was used to generalise the data to linear scale (Schäfer, 2020). Fixed effects (predictor variables) were bVt identification and discrimination prompts, Diverse Words identification and discrimination prompts, Directions, and language group. Participant and item were included as random effects with random intercepts (Brekelmans et al., 2020).

Model building followed a stepwise approach (Janssen, 2012)[35], beginning with a null model to establish a baseline for comparison and incrementally including predictors of interest. Models were compared using likelihood ratio tests and Akaike information criterion (AIC) as an estimate of prediction error (Verbyla, 2019). For AIC, a lower value reflects a better fit to the original data. The optimizer tool, BOBYQA (Powell, 2009), was used to decrease convergence errors. Individual predictors were summarised and compared using odds ratios, where 1 indicated no relation between the predictor and the outcome, greater than 1 specified greater odds, and less than 1 indicated lower odds. Odds ratios were accompanied by 95% confidence intervals created with the R package, sjPlot (Lüdecke et al., 2021). Multicollinearity was quantified by calculating variance inflation factor (VIF) with the R package, Performance (Lüdecke et al., 2021).

Fixed and random effects were next defined. Selection of fixed and random effects was made after consideration of mixed models designs and potential influences of the Diverse Sentence prompts. Placing too many parameters (fixed or random effects) in a mixed model design may lead to overparameterisation or overfitting, where more parameters are employed than can be

[35] An alternative approach, a maximal model where predictors are neither included nor excluded on the basis of significance, argues against stepwise inclusion of variables. For the current study, factors were added incrementally not only because such an approach is found in research methods texts (Janssen, 2012) and relevant literature (Xu & Lee, 2018), but because many of the predictors were related (e.g., vowel duration and word syllable count), and thus would likely "compete" to explain variability. This could yield non-significant findings for otherwise significant factors. A stepwise approach, then, permitted an exploration into the relative effects of each predictor without inadvertently losing nuanced differences.

supported by the data. Practical consequences of overfitting include creating a model which is incomputable, or modelling noise rather than the signal, and thus developing a model which cannot be generalised beyond the dataset. Conversely, underfitting the data with too few parameters may insufficiently model the signal, similarly compromising generalisability. Yang et al. (2020) explain the balance between over- and underfitting data, "underfitting the covariance structure can lead to bias in the estimated variance of the fixed effects, and overfitting could lead to the random effect covariance matrix close to singularity…inclusion of redundant covariates leads to increased prediction error" (p. 228).

As the present investigation was to identify the relative merit of using bVt prompts to predict correct responses with Diverse Sentences, the primary fixed factors were known (i.e., performance with each prompt type). What was unclear was which additional factors would significantly affect perception of the target vowels in sentences. Perception of target vowels or words in sentences may be influenced by factors such as sentence length (Holt & Wade, 2004), number of syllables (Spoehr & Smith, 1973), vowel duration (Kondaurova & Francis, 2008), position of the target word in the sentence (Marslen-Wilson & Tyler, 1975), frequency of each word (Broadbent, 1967), and phonological context. L1 listeners are familiar with perceiving contrastive phones in everyday usage—regularly displaying perceptual constancy—yet the extent to which L2 listeners are robust to such sentential and phonological diversity is uncertain. Including each of the above as fixed effects alongside prompt type predictors may lead to overparameterising. They additionally would not all suit random effects for the same reason (overfitting), but also because random effects require at least 5 levels (e.g., serial position in a sentence is three levels: initial, medial, final).

Instead of developing an unnecessarily cumbersome model with both prompt types and potential auxiliary factors, the two were split into different foci: (1) prompt types and (2) additional complexity which may help provide a better model fit. To account for the variability in each item of

Diverse Sentences, a random effect was used for Item. The specific elements which constitute complexity within the sentence prompts were examined separately in Q1e.

Q1e. To what extent does item-level complexity affect listener performance with Diverse Sentences? Or, to what extent do additional factors influence outcomes in perceiving target words in Diverse Sentence?

With the scarcity of vowel perception studies which employ connected speech prompts, the relative effects of sentence- and word-related variables were unknown. Such elements reflect confounding variables which are avoided by using fixed frame prompts for assessing vowel perception. Having documented the number of words and syllables in each sentence, vowel length of target vowels in each word[36], frequency of each word, and difference between grade levels of words in a minimal pair, these were used to predict correct responses in Diverse Sentences.

Item-level complexity was explored using generalised linear mixed models (GLMM). The GLMMs were constructed as identified in Q1d, and target vowel pairs were investigated separately. The outcome variable was obtaining a correct response with Diverse Sentences. Fixed effects were sentence length, sentence syllable count, target word syllable count, target word position, target word frequency, difference in grade level between minimally paired words, and vowel duration in connected speech. Given the target language groups documented tendency to use duration as a cue (for /i, ɪ/), and that there was a measured difference in duration between target vowels, vowel duration and number of syllables[37] in a target word were expected to be significant predictors.

Building from Pilot findings, syllables were also matched for phonological environments. Minimal pairs leave-live, feel-fill, and pet-pat had disyllabic cognates in lever-liver, feeling-filling, and petter-patter.

[36] Vowel length was subsumed in the description of sentence- and word-related variables because they modulate vowel compression.

[37] for the same reason. As noted previously, there is an indirect relationship between the number of vowels in a word and the duration of the vowels. This is termed, “vowel compression”.

RQ2. How do L2 participants' subjective experiences (as identified through an open-ended question) with diverse word and sentence prompts compare with their experience with bVt prompts? An open-ended post-experiment question was analysed to uncover salient cognitions participants held regarding specific prompt types. These were analysed qualitatively, as previously described in the section, "Open-ended question". Intercoder reliability was conducted by merging the two coder projects (Coder 2 and my own) into one NVivo file, then using "Coding Comparison" query to calculate the individual percent agreement and kappa coefficient for each code. Aggregated totals were manually calculated. These results were followed by interpretations of the findings.

Results

RQ1. Compared to bVt, to what extent are phonologically and sententially diverse prompts suitable for assessing vowel perception?

Q1a. Establishing the relative reliability of each prompt type through Cronbach's alpha was an initial, but important step in quantitatively comparing prompts. Table 13 displays Cronbach's alpha for each prompt type, with prompts adjusted to contain an equal number of items.

Table 4. Adjusted reliability (α) comparisons across prompt types (32 items, n=38)

Vowel Pair	bVt Oddity	Diverse Words Oddity	bVt Identification	Diverse Words Identification	Directions	Diverse Sentences
/i, ɪ/	.91	.78	-	.65	.51 (16 items)	.62
/ɛ, æ/	.93	.67	-	.74	.7	.73

Note. bVt Identification prompts were excluded as they were incompatible with α due to item repetition.

As shown in Table 13, bVt prompts were the most internally consistent. However, adjusted reliability for Diverse Words and Diverse Sentence prompts was moderate to strong. This indicated generally efficacious internal consistency for Diverse Words and Diverse Sentences (Taber, 2017). The Directions task was expected to yield similar results to Diverse Words Identification and thus slightly underperformed according to expectations (revisited in Discussion). A starkly reduced alpha value for /i, ɪ/ Directions can be explained by the comparatively few items in relation to other prompt types. Using the $\alpha \geq .7$ criterion established previously, internal consistency, at least in part, appears to be mitigated by vowel pair. For /ɛ, æ/, the more phonologically diverse prompts yield “acceptable” reliability, while for /i, ɪ/ results fall below the previously set criterion for acceptability.

Note that the full set of items (not adjusted downward for comparison with bVt) resulted in reliability indices for Diverse Words Identification of .81 /i, ɪ/, .82 /ɛ, æ/; for Directions it was .51 /i, ɪ/ (as reported above for 16 items), .78 /ɛ, æ/; and for Diverse Sentences, .81 /i, ɪ/ and .83 /ɛ, æ/.

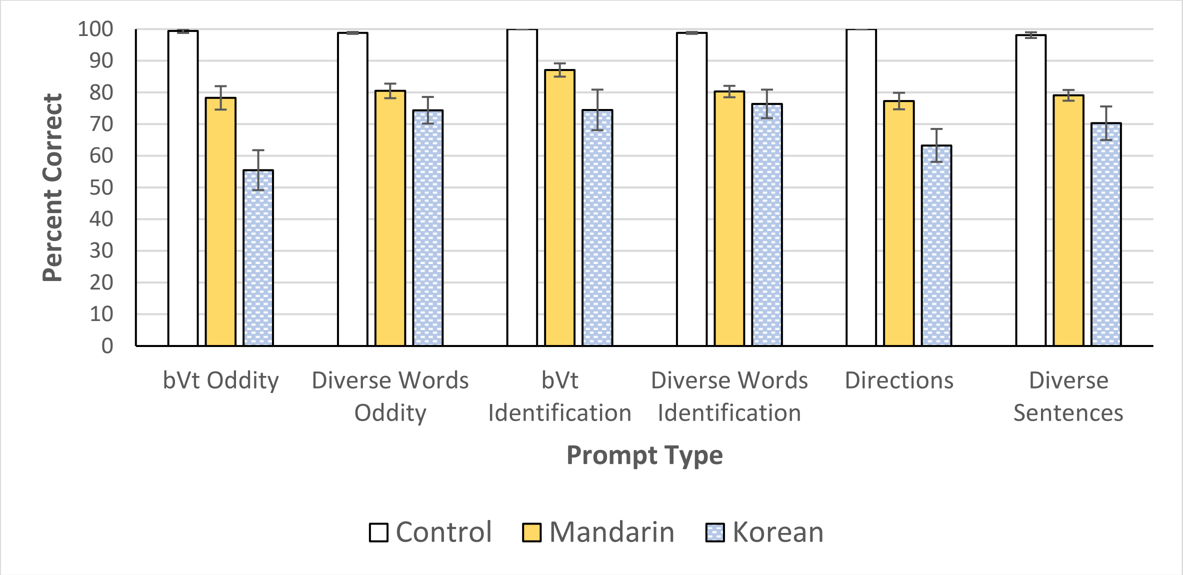
Having confirmed the comparative strength of bVt and the relative reliability of each prompt type, participant performance with each was explored to investigate how well data matched predictions for bVt prompts and the more diverse prompt types.

Q1b.

To support inferences which might be made about diverse prompt types compared to bVt, it was necessary to identify how well performance matched previous research. Recall that the control was expected to perform near ceiling levels with both vowel pairs regardless of prompt type, whereas both L2 groups were expected to perform best with /i, ɪ/ and less well with /ɛ, æ/. The Mandarin group was expected to have scores equal to or greater than the Korean group.

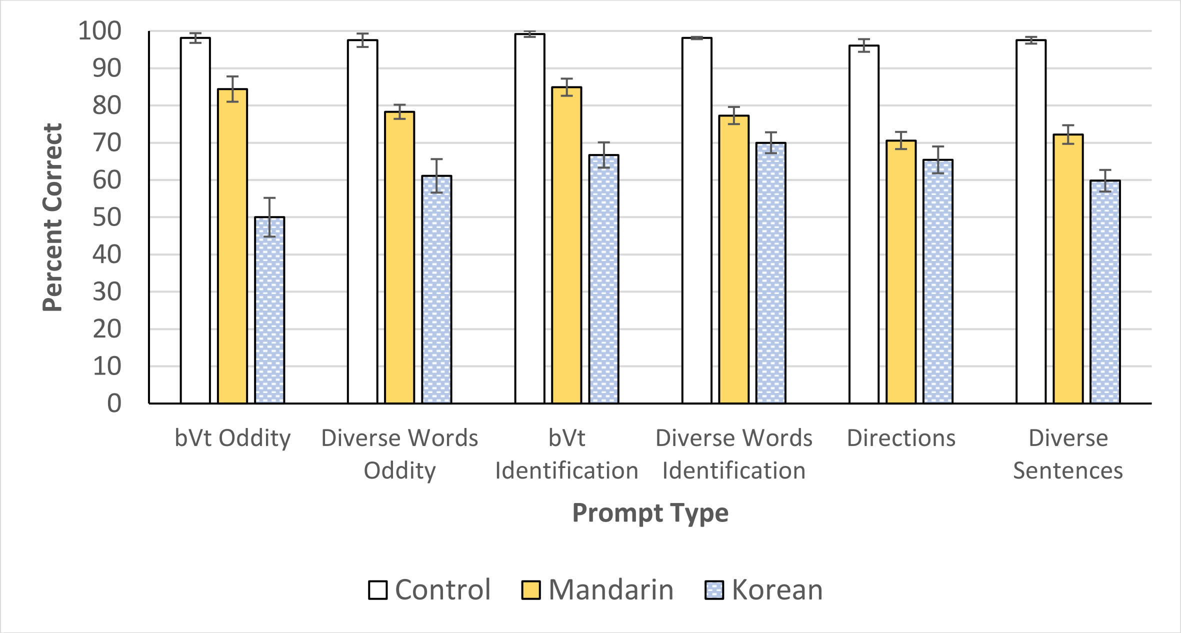
Results were generally congruent with predictions, within and between groups. The control performed at ceiling, followed by Mandarin and then Korean. Figure 3 and Figure 4 summarise descriptive statistics for the groups by prompt type.

Figure 3. /i, ɪ/ Mean Task Performance by Language Group



Note. Error bars display standard error. Chance performance for oddity was 25%; chance performance for identification tasks was 50%.

Figure 4. /æ, ɛ/ Mean Task Performance by Language Group



Note. Error bars display standard error. Chance performance for oddity was 25%; chance performance for identification tasks was 50%.

At the vowel pair level, the L2 groups were expected to not discriminate well between /ɛ/ and /æ/, but better between /i/ and /ɪ/. A dependent *t*-test comparing overall percent found that vowel pair differences in the Mandarin group (/ɛ, æ/ *M* = 76.7, *SD* = 10.8; /i, ɪ/ *M* = 80.2, *SD* = 10.7) were non-significant (*p* = .09); differences for the Korean group (/ɛ, æ/ *M* = 62.8, *SD* = 6.7; /i, ɪ/ *M* = 70.9, *SD* = 13.4) were statistically significant, *t*(7) = 2.73, *p* = 0.03. The non-significant finding from the Mandarin group was unexpected, but looking at performance by prompt type, bVt Oddity for the Mandarin group performed opposite to expectations, while other prompts were congruent with predictions. An exploratory two-way repeated measures ANOVA was conducted, breaking the overall performance score into discrete indices for each prompt to investigate whether there was an interaction between vowel pair and prompt type. Results from the two-way ANOVA revealed a significant interaction between vowel pair and prompt type, *F*(1,29) = 18.47, *p* < .001, η^2 = .39. The displayed partial eta squared suggests a medium-large effect for the interaction (Cohen, 1988), meaning the magnitude of vowel pair performance is mitigated by, or dependent upon, prompt type.

As bVt Oddity returned unexpected results, the question was why this had occurred. One initial explanation was that the studies cited to inform predictions used three rather than four interval oddity tasks. It could not be that the extra stimulus confused participants or provided a cognitive burden because performance *improved* for the vowel pair that was predicted to be challenging. Confusion or memory load would be expected to exacerbate poor scores. Further, the control trials revealed a mean percent score of 97.4 (*SD* = 9.7) for /ɛ, æ/, and 96.7 (*SD* = 11.9) for /i, ɪ/. Instead, the extra stimuli appeared to have a facilitatory effect, allowing participants to attentively grasp distinctions otherwise unknown, or it may simply be an artifact of the experiment.

It is unclear without further investigation why this would aid the Mandarin group and not the Korean group, or whether this finding is replicable.

Similar to the Mandarin group, the Korean group appeared to have a prompt with results contra to predictions. The Directions task singularly displayed lower /i, ɪ/ scores than /ɛ, æ/. However, a dependent *t*-test found the difference between groups to be non-significant.

Consolidating these findings, the Diverse Words and Diverse Sentences prompts performed as expected, as did bVt Identification. The single statistically significant finding contrary to predictions was the bVt Oddity prompt, where Mandarin outperformed expectations for /ɛ, æ/. Together, this suggests efficacy for bVt Identification, Diverse Sentences, and Diverse Words prompts.

Q1c. Between-prompt complexity was explored through percent correct and *d'*. Table 5 summarises Mandarin and Korean L1 groups' percent correct and d' for each prompt by vowel. Higher *d*' indicates participants displayed greater sensitivity to the distinctions in the vowels. Recall that a linear decrease in *d'* was expected as complexity increased in prompts (Table 5 organises prompts from least complex on the left [bVt] to most complex on the right [Diverse Sentences]). Further, as the L2 groups were expected to perform better with /i, ɪ/ than /ɛ, æ/, *d'* should be higher for /i, ɪ/ than /ɛ, æ/.

Table 5. Identification accuracy of vowels by prompt type (n =38)

	Prompt type							
	bVt Identification		Diverse Words Identification		Directions		Diverse Sentences	
Vowel	%	*d'*	%	*d'*	%	*d'*	%	*d'*
/i/	89	2.27	82	1.64	77	1.31	78	1.50
/ɪ/	88	2.27	81	1.64	78	1.31	81	1.50
/ɛ/	84	1.96	77	1.40	72	1.02	75	1.03
/æ/	87	1.96	80	1.40	73	1.02	71	1.03

Note. Percent of correct responses (%) and scores converted to d prime (d') are provided.

Identification and oddity tasks had separate chance scores and were considered separately. Results for identification tasks generally supported expectations across prompts and vowels. One-way repeated measures ANOVAs revealed a statistically significant difference in prompt type performance for both vowel pairs. Mauchly's test of sphericity was significant ($p < .05$) for both vowel pair analyses, and a Greenhouse-Geisser correction was used. For /ɛ, æ/, $F(2.5, 90.8) = 8.01$, $p < .001$, $\eta^2 = .18$. Post hoc testing confirmed a significant ($p < .001$) linear decline in correct scores for bVt Identification ($M = 81.0$, $SD = 13.9$), Diverse Words ($M = 75.8$, $SD = 12.1$) and Directions ($M = 69.5$, $SD = 12.1$). The difference between Directions and Diverse Sentences ($M = 69.6$, $SD = 13.8$), the two connected speech prompt types, was non-significant ($p > .05$). Similar results were found for /i, ɪ/, $F(3.1, 113.8) = 8.39$, $p < .001$, $\eta^2 = .19$. Post hoc testing showed that the linear decrease between bVt Identification ($M = 84.4$, $SD = 14.0$), Diverse Words Identification ($M = 79.4$, $SD = 10.7$), and Directions ($M = 74.3$, $SD = 15.4$) was significant ($p < .01$). Mirroring /ɛ, æ/, the difference between /i, ɪ/ Directions and Diverse Sentences ($M = 77.2$, $SD = 11.0$) was non-significant ($p > .05$).

Results were mixed for the oddity tasks. For /ɛ, æ/, the difference between the two oddity tasks, bVt Oddity ($M = 77.1$, $SD = 22.6$) and Diverse Words Oddity ($M = 74.7$, $SD = 12.9$), was significant ($p < .05$). The difference between /i, ɪ/ was also statistically significant ($p < .05$); however, it was opposite to expectations as bVt Oddity ($M = 73.5$, $SD = 21.9$) resulted in lower scores than the more phonologically diverse prompt type, Diverse Words Oddity ($M = 79.2$, $SD = 12.7$).

D' largely reflected predictions, with the exception of Directions. *D'* was lower for Directions than for Diverse Sentences, compromising the prediction of linearity across prompts. Though Diverse Sentences was the most phonologically diverse prompt type, participants were more sensitive to the differences in target vowels in Diverse Sentences compared to Directions.

Though the Directions prompt type was designed to provide a relatively easy connected speech task with uniform syntax, by *d'*, it was found to be the most difficult. It was postulated that the Directions prompts contained streets which were unfamiliar to participants, and that may have resulted in poorer performance compared to more familiar words. Number of syllables was also considered. Out of the nine pairs of prompts, only two were monosyllabic. Six were disyllabic and one was trisyllabic. This could have influenced perception through vowel contraction (where vowel duration is inversely related to the number of syllables in a word) or simply by distracting the participant with additional acoustic information. This will be explored further with the mixed model results.

Despite the unforeseen difficulty of Directions, sensitivity indices across all prompt types were within the prescribed desired range of 0.5-2.5, the boundaries above chance and below perfect scores (MacMillan & Creelman, 2005). Statistically, the easiest prompt types were not too easy, nor were the most difficult too hard.

This section reported effects of complexity from an aggregate, prompt-level perspective. The effects of individual variables which constitute complexity (in Diverse Sentences) are later explored as part of a mixed model design (following next section). The next section introduces results from the initial mixed models, investigating the bVt prompts' predictive facility.

Q1d. Answering "to what extent do isolated bVt prompts predict Diverse Sentence performance?", a series of generalised linear mixed models were run, with results shown in Table 6 and Table 7. The bVt prompts (m1) revealed a significantly better fit than the null for both /i, ɪ/, $\chi^2(2) = 43.0$, $p < .001$, and /ɛ, æ/, $\chi^2(2) = 41.1$, $p < .001$. The next step was to compare bVt prompt fit with Diverse Words (m2), the isolated counterpart to bVt. Though the pairs of variables could not be directly compared with each other using likelihood ratios—there were 0 degrees of freedom between m1 and m2—Diverse Words prompts displayed a lower AIC than the bVt prompts when both were compared to the null for /i, ɪ/ (bVt AIC = 2308; Diverse Words AIC = 2390) and /ɛ, æ/ (bVt AIC = 2510; Diverse Words AIC = 2498), suggesting a better fit for Diverse Words. Additionally, all four prompts together (m3) fit the data significantly better than bVt prompts alone (/i, ɪ/ $\chi^2[2] = 12.5$, $p < .001$; /ɛ, æ/ $\chi^2[2] = 13.1$, $p < .01$), but did not fit the data significantly better than Diverse Words prompts alone (/i, ɪ/ $\chi^2[2] = 1.7$, $p > .05$; /ɛ, æ/ $\chi^2[2] = 1.08$, $p > .05$). Stated alternatively, including bVt in a model with Diverse Words did not materially alter findings. For this study, Diverse Words were a better fit for the data and more effective predictors than the bVt prompts.

Odds ratios (*OR*) enabled a comparison of individual predictors with confidence intervals (*CI*) determining significance. Diverse Words Identification was the strongest single predictor of Diverse Sentence performance, as higher performance in Diverse Words Identification associated with higher performance in Diverse Sentences (/i, ɪ/ *OR* = 1.04, *95% CI* = [1.01, 1.05]; /ɛ, æ/ *OR* = 1.04, *95% CI* = [1.02, 1.07]).

A maximal model with all predictors yielded non-significant predictors. This may be explained by multicollinearity between prompt types—as expected given they are attempting to measure the same thing in different ways. When combined into a maximal model, the collinear variables competed to explain the data, negating their respective significance (Frazier & Tix, 2004). Presumably, there is a moderating third factor which is influencing results for each prompt type. This third factor is assumed to be perceptual constancy, measured at different levels by the different prompt types.

Directions was also included as a predictor (m3), but was automatically dropped from the full model for being “rank deficient”. Rank deficiency indicates that one variable (the one dropped from analysis) contained the same information as a variable already present. Models with Directions were always rank deficient when accompanied by Diverse Words Identification. Replacing Diverse Words with Directions yielded the same outputs; presumably these explanatory variables provided the same information.

Revisiting Q1b, language group was included, showing a non-significant difference in L2 group performance as indicated by confidence intervals for both vowel pairs (/i, ɪ/ *OR* = 0.81, *95% CI* = [0.57, 1.14]; *OR* = 1.07, *95% CI* = [0.62, 1.83]). The control, as expected, displayed significantly better results than Mandarin, and thus Korean[38] (/i, ɪ/ *OR* = 5.92, *95% CI* = [2.47, 14.20]; *OR* = 6.20, *95% CI* = [2.50, 15.34]. Such results match previously indicated research.

[38] For multilevel factors, odds ratio outputs for generalised linear mixed models select a comparison group based on alphabetic or numeric order. The Mandarin group was arbitrarily coded as 1 (Korean = 2, control = 3) in the dataset, making it the default “level” to which the other groups were compared in the analysis. Therefore the control group was directly compared to the Mandarin group, but only indirectly (through Mandarin) to the Korean group. To compare each level with the comparison group, an odds ratio of 1 is automatically assigned to the default group. For this analysis, less than one indicates poorer performance than Mandarin, while greater than 1 indicates better performance than Mandarin.

Table 6. Experiment 1 /i, ɪ/ model comparison of prompt and language predictors for Diverse Sentences

					LRT		
Model	Fixed effect	Deviance	*df*	AIC	comparison	X^2(*df*)	*p*
m0	-	2441	3	2447			
m1	bVt Odd, bVt ID	2398	5	2408	m0-m1	43.0(2)	<.001***
m2	Diverse Words Odd, Diverse Words ID	2380	5	2390	m0-m2	61.0(2)	<.001***
m3	bVt Odd, bVt ID, Diverse Words Odd, Diverse Words ID	2378	7	2392	m1-m3	12.5(2)	<.001***
					m2-m3	1.7(2)	>.05
Adjusted							
m0	-	1845	3	1851			
m1	bVt Odd, bVt ID	1809	5	1819	m0-m1	36.1(2)	<.001***
m2	Diverse Words Odd, Diverse Words ID	1791	5	1801	m0-m2	53.9(2)	<.001***
m3	bVt Odd, bVt ID, Diverse Words Odd, Diverse Words ID	1790	7	1804	m1-m3	28.9(3)	<.001***
					m2-m3	1.6(2)	>.05

Note: Table shows model with all data (top) and an adjusted model with overlapping words removed (bottom). Random effects included Participant and Item for all models. AIC = Akaike information criterion. LRT = Likelihood ratio test. Odd = oddity task. ID = identification task.

Table 7. Experiment 1 /ɛ, æ/ model comparison of prompt and language predictors for Diverse Sentences

Model	Fixed effect	Deviance	*df*	AIC	LRT comparison	X^2(*df*)	*p*
m0	-	2541	3	2547			
m1	bVt Odd, bVt ID	2500	5	2510	m0-m1	41.1(2)	<.001***
m2	Diverse Words Odd, Diverse Words ID	2488	5	2498	m0-m2	53.2(2)	<.001***
m3	bVt Odd, bVt ID, Diverse Words Odd, Diverse Words ID	2485	7	2499	m1-m3	21.6(2)	<.001***
					m2-m3	9.4(2)	>.05

Note: Random effects included Participant and Item for all models. AIC = Akaike information criterion. LRT = Likelihood ratio test. Odd = oddity task. ID = identification task.

Q1e. Exploring relative effects of sentence- and word-level variables, a series of generalised linear mixed models was constructed. Results are summarised for /i, ɪ/ in Table 8. Experiment 1 /i, ɪ/ Model comparison for auxiliary variables which may help predict Diverse Sentencesand /ɛ, æ/ in Table 9.

For /i, ɪ/, there was a significant effect for sentence length (*OR* = 1.3, 95% *CI* = [1.10, 1.66]), though not in the direction anticipated; longer sentences led to better odds of accurately identifying target words. Note that sentences used in the experiment were restricted in length (*M* = 5.8 words, *SD* = 1.6), and manifestly there would be a cut-off point where the opposite becomes true (where a longer sentence would lead to poorer perception). If replicable, a curvilinear relationship may be assumed. It was posited that the longer sentences, to the extent employed in Experiment 1, provided additional processing time for participants.

The same sentence duration effect was not found for /ɛ, æ/. Recalling that participants were expected to perform well with /i, ɪ/ compared to /ɛ, æ/, the stronger working perception may have afforded listeners the ability to make use of the additional time. With /ɛ, æ/, the additional time was irrelevant as participants did not have strong enough discrimination to be able to make use of it. The relative impact of sentence length compared to other predictors is displayed in Table .

Table 8. Experiment 1 /i, ɪ/ Model comparison for auxiliary variables which may help predict Diverse Sentences

Model	Fixed effect	Deviance	*df*	AIC	LRT comparison	X^2(*df*)	*p*
m0	-	2441	3	2447			
m1	Word frequency band	2434	9	2452	m0-m1	6.6(7)	>.05
m2	Word frequency band*frequency band similarity[1]	2427	13	2452	m0-m2	14.1(3)	>.05
m3	Sentence length	2432	4	2440	m0-m3	8.3(1)	<.01**
m4	Sentence length, sentence syllable count	2430	5	2440	m3-m4	2.9(1)	>.05
m5	Sentence length, sentence syllable count, multisyllabic	2429	6	2441	m3-m5	3.9(2)	>.05
m6	Sentence length, sentence syllable count, multisyllabic, vowel /ɪ/	2428	7	2442	m3-m6	5(3)	>.05
m7	**Sentence length, sentence syllable count, multisyllabic*vowel /ɪ/**	**2418**	**8**	**2434**	**m3-m7**	**14.2(4)**	**<.01****

Note: Random effects included Participant and Item for all models. AIC = Akaike information criterion. LRT = Likelihood ratio test. Odd = oddity task.

ID = identification task. Bold text reflects the best fit model.

[1] No model which included lexical frequency was significant. Frequency was omitted from subsequent models.

Table 9. Experiment 1 /ɛ, æ/ Model comparison for auxiliary variables which may help predict Diverse Sentences

Model	Fixed effect	Deviance	*df*	AIC	LRT comparison	$X^2(df)$	*p*
m0	-	2541	3	2547			
m1	**Word frequency band**	**2514**	**8**	**2530**	**m0-m1**	**27.1 (1)**	**<.001*****
m2	Word frequency band*frequency band similarity	2510	12	2534	m2-m1	3.9(4)	>.05
m3	Sentence length	2541	4	2549	m0-m3	2.7(1)	>.05
m4	Sentence length, sentence syllable count	2539	5	2549	m0-m4	2.7(2)	>.05
m5	Sentence length, sentence syllable count, multisyllabic	2535	6	2547	m0-m5	4.9(3)	>.05
m6	Sentence length, sentence syllable count, multisyllabic, vowel /æ/	2536	7	2550	m0-m6	5.7(4)	>.05
m7	Sentence length, sentence syllable count, multisyllabic*vowel /æ/	2534	8	2550	m0-m7	1.7(1)	>.05

Note: Random effects included Participant and Item for all models. AIC = Akaike information criterion. LRT = Likelihood ratio test. Odd = oddity task.

ID = identification task. Bold text reflects the best fit model.

The pilot indicated a potential interaction between vowel and syllable type, and thus these variables were explored with Diverse Sentences. As shown in Table , for /i, ɪ/, multisyllabic words had lower odds (indicated by *OR* < 1) of being answered correctly compared to monosyllabic words (*OR* = .32, *CI* = [.13, .75]), while target words with /ɪ/ were nominally, but not statistically significantly, more challenging compared to /i/. Yet when the word was both multisyllabic and had /ɪ/ as the target vowel, the odds of obtaining a correct response were considerably higher (4.69 times) than when the word is multisyllabic and /i/.

Table 19. /i, ɪ/ optimal model (M7) outputs

Predictors	Odds Ratios	Confidence intervals
(Intercept)	9.52	2.76 – 32.87***
Sentence length	1.3	1.10 – 1.66**
Sentence syllable count	0.96	0.81 – 1.15
Multisyllabic	0.32	0.13 – 0.75**
Vowel /ɪ/	0.83	0.51 – 1.35
Multisyllabic * vowel /ɪ/	4.69	1.81 – 12.12***
Random Effects		
σ^2	3.29	
$\tau_{00\ \text{item}}$	0.53	
$\tau_{00\ \text{participant}}$	0.96	
ICC	0.31	
$N_{\text{participant}}$	43	
N_{item}	64	
Observations	2752	
Marginal R^2 / Conditional R^2	0.058 / 0.352	

Note. Table shows predictors of correct responses for Diverse Sentence prompts.

Frequency was not a significant predictor for /i, ɪ/ but was for /ɛ, æ/. For /ɛ, æ/, the most common word frequency (level K1) offered higher odds of obtaining a correct response. Results for word frequency were non-linear, however. Whereas levels K3, K4 and K17 were more difficult than

K1, K17 was easier than K3 and K4. With the non-significant findings of /i, ɪ/ and the significant, but non-linear findings of /ɛ, æ/, no firm assertions may be made. Given the expected effect of word frequency and the mixed results found herein, no definitive claims will be made. Instead, refinement of the instrumentation is suggested to better target participants' idiosyncratic vocabularies (see Discussion).

Absolute vowel duration was not found to be a significant predictor for either vowel pair.

RQ2. How do L2 participants' subjective experiences (as identified through an open-ended question) with diverse word and sentence prompts compare with their experience with bVt prompts?

To answer this question, I begin by reporting results from the intercoder reliability analysis, summarised in Table 10.

Table 10. Kappa agreement coefficients and percentage of absolute agreement for Experiment 1 codes

Code	Kappa coefficient	Percentage of absolute agreement	Total occurrences
Attention\general	0.6354	95.75	8
Attention\confusion	0.6113	89.4	16
Attention\fatigue	0.7767	99.5	6
Emotion\negative affect[1]	0.7148	99.1	3
Emotion\positive affect[1]	0.3467	97.55	3
General cognition	0.3504	90.68	12
Perceived difficulty	0.8443	92.91	45
Strategies	0.7649	95.67	8
Prompt\bVt	0.6711	90.46	29
Prompt\diverse words	0.5821	88.72	26
Prompt\directions	0.7267	93.01	22
Prompt\diverse sentences	0.8075	94.06	25
Average	0.6527	93.90	203

[1]Emotion (Positive and Negative Affect) was regrouped as General Cognition

Overall results fit Landis and Koch's (1977) criteria for substantial intercoder reliability. The lower individual kappa agreement for Positive Affect and General Cognition compared to the other codes can be explained by the scarcity of references (Positive Affect) or discretionary nature of the code (General Cognition).

Responding now to the question, “What were the salient cognitive effects of employing fixed and diverse listening prompts for assessing L2 vowel perception?”, I found a single overarching theme: “cognition during the experiment”. I defined *cognition* as thoughts or thought processes participants had regarding specific prompt types, and I was primarily interested in how thoughts about bVt prompts compared and contrasted with other prompt types. I have used pseudonyms to report participant quotations.

The predominant cognition that participants reported related to perceived prompt difficulty. There was a mixture of difficulty responses between bVt and sentence prompts, with some participants labelling bVt prompts as difficult and sentence prompts as easy, while others considered sentence prompts easier and bVt prompts more difficult.

Several participants felt that hearing words in a sentence provided an advantage compared to bVt prompts and isolated words, as indicated by Ketan, “We can choose the words according to the context, what the words mean in the sentences, which would be different from the former parts”. Beyond strategically using context to decipher meaning, one participant mentioned duration as a potential advantage, “when put into the sentences, some sounds are more easier to be recognized like meals and mills, maybe it is because i have some more time to prepare” (Shalini). Shalini’s sentiment here directly supports the preliminary hypothesis posed during the mixed models analysis, that longer sentences may provide addition processing time for participants.

Lorena offered another potential perceptual advantage, that sentences offered diversity that was absent in isolated prompts: “bvt-Discrimination, Diverse Words-Discrimination, bvt-Identification ARE MORE DIFFICULT THAN Diverse Sentences and Directions BECAUSE THEY LOOK SIMILAR”. Here, the repetitive nature of single, isolated word prompts increased task difficulty.

It appears that for some participants, sentences are better suited to retain attention. Though the sentence-based prompts held the most items and were the longest in duration (with Diverse

Sentences the longest), the only participant responses which explicitly identified length and tedium of a specific prompt type were directed at isolated words. As Arpan states, the bVt prompts were "not very challenging but take effort when I am not 100% attentive". Similarly, another participant explained to me after the experiment that she went on "autopilot" during the bVt identification task and occasionally misclicked[39]. This was not specific to the bVt context, however, as Monica wrote of Diverse Words, "Overall easy, but when I hear these words reappearing over and over again, I felt less concentrated and I misclicked one of the choices".

Views of sentence prompt efficacy were not universally or unreservedly held. Carol noted that the utility of context may be contingent upon familiarity, "If I can understand the word in a sentence, it's much easier to guess". (Familiarity will be discussed separately.) Richard felt that using context to augment listening in the study was dubious,

> Examples in Diverse Sentences can be misleading to me. Because usually i identify differences words and meanings based on contexts (meaning i sometimes tend to choose the word that i feel the right in the sentences inevitably rather than fully concentrating on what i've heard).

A final perceived disadvantage exclusive to sentence prompts—distraction caused by word or sentence level features—requires additional consideration. As Judith typed, "In sentence, the stress on other words may lead me to ignore the target pronunciation". Here, the participant infers sentence prompt inefficacy compared to bVt prompts, that embedding words in sentences is distracting for vowel perception tasks. This statement targets the heart of the present research. Are features which are inherent in sentences (e.g., sentence stress, context) distractions or construct relevant variables? Is it sufficient to identify effective perception of a vowel in a single environment or context (i.e., in isolation with invariable consonantal neighbours), or is it meaningful to uncover

[39] The participant was lamenting she was not able to change responses. It might be worth including such an option for isolated word tasks.

whether the participant’s perception extends beyond isolation such that their vowel perception demonstrates constancy[40]?

Beyond subjective and theoretical positioning, Judith’s claim speaks to the task design as much as the task objective. The participant is aware that she did not need to listen to the sentence, only the labelled words. As indicated in the quantitative results, the labelled buttons on screen permitted participants to ignore the sentence, reducing the sentence context to noise and making the sentence perception mirror isolated words (with sentence-related durational and phonological influences). The sentence as a “distraction” is a disadvantage to listening for a specific vowel, but an advantage for assessment if assessing one’s ability to perceive speech beyond an isolated word.

Other noted features of difficulty were not specific to sentences, but diverse prompts (i.e., Diverse Words and Directions) more generally. Two participants indicated that lack of familiarity made perceiving locations difficult, and another, that having multiple syllables exacerbated task difficulty.

> Some words like "Sheep Lane" "Granville Place" is quite demanding, as there are multiple syllables in the words and it took me some time to work out which vowel the discrimination lie in and the stress and intonation of the words kind of distract me. (Michael)

Such reports are consistent with the initial reasoning behind low scores reported for the Directions Task. Directions contained multisyllabic words, shown to correspond with lower scores than monosyllabic words, and names of places unfamiliar to participants. Descriptors for the task typically indicated difficulty, ranging from “a little hard to distinguish”, to “couldn’t distinguish”, “quite demanding”, “really confusing”, and “most difficult”.

Participant cognitions mentioned here were helpful in providing a more well-rounded approach to the data, offering a glimpse of participant experience with each prompt type. Quantitative and qualitative results have now been displayed for each research question. The

[40] Constancy is the ability to perceive an object as the same object despite contextual change.

ensuing section consolidates and discusses these findings in the context of the research aims and established literature.

Discussion

This study investigated the relative effects of employing prompt types of varying degrees of complexity for assessing vowel perception. The purpose was twofold- first to identify the potential suitability of diverse prompt types, and second to uncover how well employing a prompt type with limited complexity (bVT) would enable researchers to infer participants' connected speech (Diverse Sentences) performance. This was done with the overarching intent to inform practical assessment practices. Addressing the first purpose, quantitative and qualitative data were obtained.

As may be expected, greater complexity resulted in a more difficult prompt, evidenced by percent correct and *d'*. Notably, however, increased complexity did not prevent prompts from being effective. The most complex prompt type, Diverse Sentences, displayed acceptably strong internal consistency (even when numbers were reduced for cross-prompt comparison), kept within prescribed ranges for *d'*, and strictly adhered to PAM-L2 predictions. Such adherence may be contrasted with bVt Oddity, where the Mandarin group unexpectedly performed better with the mid-low vowel pair than with the high vowel pair.

With the understanding that bVt prompts have been employed canonically to prevent extraneous variables from influencing results, this study investigated a sampling of variables which would be controlled for in a bVt study. A significant effect was found for sentence length, with greater length resulting in greater performance for the /i, ɪ/ contrasts. This was reasoned to be a consequence of additional processing time, a notion anecdotally supported by Shalini, the participant who believed sentences provided additional time to prepare for the contrasts. The same effect was not found for /ɛ, æ/, and it was speculated that due to SC assimilation, participants did not have a strong enough fundamental contrast between /ɛ, æ/ to make use of the additional processing time. Based on sentence duration results, preliminary hypotheses for future research may map to the PAM-L2 hierarchy as descriptors:

> TC (readily contrasts phones, no need to rely on additional cues in clear speech conditions) > CG (can make use of additional processing time to help discriminate between target contrasts) > SC (not able to make use of sentential cues to identify target contrasts).

Targeted work exploring how sentential features impact L2 perception at various levels of assimilation would be a valuable contribution to perceptual assimilation models, though it is beyond the scope of the current research.

Other variables used to explore the effects of complexity within Diverse Sentences revealed non-significant results. The lack of significance was of interest, particularly for vowel duration (and number of syllables in a target word for similar reasons). As discussed in Methods, L2 participants have been found to prioritise temporal cues over spectral cues for target vowels, and the target words did contain systematic durational differences (see Target vowel duration.). The non-significant finding may be explained by the varied durations of each vowel and each vowel token[41]. Target vowels in multisyllabic words and connected speech were generally shorter than their monosyllabic, isolated word counterparts, and vowel durations were modulated by their phonetic environments, such as when they preceded voiced and voiceless consonants, making them longer or shorter, respectively. This attenuated the ability for listeners to use duration as a cue for vowel identification. Whether this (and speech signal diversity in general) is relevant to the construct of listening perception will be revisited.

Prompt type suitability was also filtered through the lens of participant experience. Feedback indicated that for some participants, isolated prompts were associated with repetition and tedium, occasionally resulting in "misclicks". A definitive claim that using sentences in speech perception tasks reliably promotes participant attention or mitigates fatigue would require more empirical support, but some participants indicated that the sentence prompts were helpful by way

[41] *Vowel token* was referenced separately from *vowel* as each instance of each vowel was a unique (allophonic) representation of the vowel.

of context and, as noted, time to prepare. Considered together with quantitative findings, the diverse prompts, especially Diverse Sentences, compared favourably with bVt, indicating an efficacious option for assessing vowel perception in L2 speakers.

With a provisional establishment of Diverse Sentences as a viable prompt type, the emergent question was how well bVt prompts predicted perception of target vowels in Diverse Sentences. The bVt oddity and identification prompts were found to model the data better than a null model, but not as well as the Diverse Words prompt types. Diverse Words Identification was the best model fit and strongest predictor of Diverse Sentence performance, yet the effect was small (shown by the 1.03 odds ratio), suggesting that generalising perception in sentential contexts from isolated contexts should be done with caution. To identify performance in sentential contexts, sentential prompts should be employed. If isolated words are necessary and generalisation to connected speech is, Diverse Words are preferrable to a single CVC frame.

A sample of factors which are controlled for in isolated bVt prompts but are present in sentence prompts and may affect performance were examined. Sentence length was found to be a significant predictor for /i, ɪ/ (odds ratio of 1.3), but not /ɛ, æ/. Interpreted in the framework of PAM-L2, such a result may be due to CG offering a working level of differentiation that may be facilitated with additional processing time given by sentences, whereas SC does not distinguish between the phones well enough to effectively use such cues. Though intriguing, this may be a function of the task design as it is unclear additional processing time would exist without providing written text of the minimal pair that participants needed to choose between. Further inquiry is required.

Word frequency was expected to have an effect in accurately perceiving target words in sentences, yet results were mixed. It may be that limitations in the experiment's design meant that effects went undetected or improperly detected. The measure used to identify familiarity—frequency in the BNC—may not have adequately described individual participants. The BNC was used as a rough predictive tool to aid design decisions, but a more refined measure that accounts for individual differences in word familiarity is needed to appropriately make claims about the impact of words in sentences for assessing vowel perception.

Despite mixed results with word frequency, word familiarity was still thought to have been responsible for performance in connected speech, especially the sub-optimal prompt performance in Directions. Numerous participants mentioned familiarity and frustration in connection with the Directions prompts. Disposition and mood have been found to impact performance (Uddenberg & Shim, 2015), and this may have helped lead to lower performance than expected.

Syllabicity and vowel type appear to interact for /i, ɪ/ vowels, impacting L2 listeners' tendencies to accurately identify the target word in Diverse Sentences. When the target was multisyllabic with /ɪ/, it was more likely to be perceived accurately than /i/. This was an interesting finding as multisyllabicity and /ɪ/ words individually led to lower performance. The interaction may be explained through L2 assimilation and vowel contraction. As vowels become shorter in words with more syllables and connected speech, L2 listeners who use duration as a primary cue to distinguish the two vowels may easily confuse the shorter /i/ for /ɪ/, while L2 listeners would more readily identify /ɪ/ accurately as it is associated with being short. Absolute duration, however, was not a significant predictor for accuracy. Similarly to word frequency, it may be explained as a problem in the appropriacy of the measure. If participants had an internal durational criterion, where high front vowels longer than a given duration are perceived as /i/, while shorter are /ɪ/, then measuring the durations to make a prediction would be misleading. This would be analogous to predicting the score of a football match by measuring the distance the ball travels in one direction or the other, but not knowing the size of the field. A more useful way of using duration as a predictor is through a general grouping of monosyllabic (a shorter field) and multisyllabic (a longer field).

Extrapolated to vowel perception assessment in general, there is practical and theoretical merit for employing either fixed or diverse prompts. Fixed prompt types enable a controlled focus of vowel perception within a limited context. Results are reliable, yet generalisability is questionable as performance with fixed frame prompts has only a minor effect in predicting performance in connected speech. If perceptual constancy—where a listener who can identify target vowels in varied contexts—is relevant to the construct of listening perception, additional diversity in listening prompts may be suitable.

The argument of which to employ for effectively assessing vowel perception, may be considered one of construct relevance (whether adding greater diversity to listening prompts helps assess vowel perception) versus construct irrelevant variance (where one is testing a different construct than intended). Here, however, I would argue that both bVt and sentential prompts *do* assess the same construct; that it may not be about construct irrelevant variance, but of how much variance is necessary to adequately model constancy. In this way, it becomes a contemplation of construct fulfilment.

This study cannot resolve such a dispute, but it can highlight the importance of its consideration. For instance, if a vowel perception identification task controls for vowel duration, spectral cues will be fronted (i.e., more prominent than duration cues), but the instrument still cannot identify whether participants are able to use relevant acoustic data to identify the vowel of interest while ignoring irrelevant data (i.e., it cannot assess perceptual constancy)[42]. Conversely, if an instrument does not control for duration but employs a restricted CVC frame, the durational cues will be pronounced, potentially leading to an assessment of temporal discrimination rather than vowel discrimination. This may help explain mixed reports in the literature (c.f. Nelson & Kang, 2015; Kim, 2010).

Experiment 1 introduced variability (phonological and sentential) to the construct of vowel perception assessment, inducing listeners to identify relevant boundaries in context (Gokgoz-Kurt & Holt, 2018) at normal speed[43]. Overall, the added variability led to more challenging tasks compared

[42] Researchers assess this, to an extent, in oddity tasks with "catch" trials. Catch trials include a full set of the same word, meaning there is no odd word in the trial. In this way, participants must identify which differences are contrastive and which are not. It would be valuable to find how well this predicts constancy in more diverse conditions.

[43] "Teacher talk" may not well reflect speech in a typical conversation (Gokgoz-Kurt & Holt, 2018, p. 305) and Talkers were encouraged to speak at a normal rate. See Recording in Methods.

to bVt prompts based on task design, and this was corroborated by learner perceptions of task difficulty. Results, however, remained preliminary.

Several methodological decisions in Experiment 1 were expected to have inflated participant scores, possibly blunting distinctions between prompt types. Participants had been directed to target words through button labels on screen, cuing their attention and permitting them to ignore the meaning of each sentence. For participants who chose to ignore the sentences, perception was akin to isolated contexts, where attention is fixated on a single word and the sentence was essentially noise. Additionally, the same-different task design likely inflated results, as there was a 50% chance of guessing correctly, which may have disproportionately affected prompt performance. Connected speech prompts yielded lower scores than bVt, suggesting greater difficulty, and consequently, these more challenging prompts may be more prone to guessing. An inflation of scores through guessing could have diminished performance distinctions between traditional prompts and the connected speech prompts. Altering response options became an immediate interest.

It was postulated that given the individual nature of language learning, a person's familiarity of a word may not be sufficiently represented by a general measure such as frequency. Further, with sentences, familiarity may be subordinate to association, as one word in a minimal pair might be more associated with the sentence even if it is relatively less familiar to the listener in isolation. To better understand sentential prompt functioning for the purpose of vowel perception task design, it would be useful to identify how familiarity interacts with participant responses.

Two additional questions thus emerged from Experiment 1: "how would participants perform on each task when they were not cued by labels?" and "what are the effects of association on perception in connected speech prompts?" A second experiment was designed to address these questions.

Experiment 2

The purpose of this research was to inform vowel perception instrument development by investigating the effects of phonological and sentential diversity in listening prompts. Experiment 2 extended findings from Experiment 1, investigated the effect of target word familiarity and association on responding to sentence prompts and reducing score inflation by using an open-ended response type (transcription). Further, it provided a deeper understanding of the participant experience with each prompt type by incorporating post-task surveys, a modified NASA Task Load Index (NASA TLX).

Methods

Recruitment and administration amidst Covid-19

Recruitment and administration were compromised by the Covid-19 pandemic. Recruitment took place in early 2020, coinciding with the mass uncertainty and rampant spread of Covid-19. Consequently, recruitment activities were switched to strictly word of mouth and posters (see Experiment 2 Appendix) were modified to include all necessary information for an online administration, including QR codes and direct links to the information sheet and the online portal to start the experiment.

Administration was shifted entirely online and documents (language background questionnaire, NASA TLX, vocabulary surveys) were modified accordingly. Initially it was possible to continue with university-based administrations, but as Covid-19 worsened and participants fled to their home countries (or made plans to do so), administrations pivoted to participants' homes before ceasing entirely. Once administrations became home-based, it was no longer possible to maintain laboratory settings and a list of self-monitoring conditions was created to promote uniformity. These conditions (headphones or earphones, a quiet room free from distraction for up to two hours, reliable wifi) were communicated in emails to potential participants and a warning was placed at the beginning of the experiment to remind participants of the requirement before beginning (see Appendix: Warning). The online platform permitted a recruitment policy for device type, and devices were restricted to computers at the exclusion of tablets and smartphones. Performance of at-home and at-university administrations was monitored and will be discussed in Methods and Results.

Participants

Talkers. The same two male and two female British English voice actors identified in Experiment 1 Methods were used here. There were additional words recorded for the Travel Agent task (discussed in Recordings).

Listeners.

Fifty-two normal hearing, adult students from a London, UK, university took part in Experiment 2 (age 18-51, *M* = 27.8 years, *SD* = 7.3). Forty-six participants were target L2 speakers (33 Mandarin L1, 13 Spanish L1), 6 were a control (English L1). The L2 group began learning English at 3-19 years (*M* = 8.6, *SD* = 4.0), had studied English an average of 14.5 years (*SD* = 5.4), and had previously demonstrated their English language proficiency in a standardised test, with an average overall IELTS score of 7.3 (*SD* = 0.6) and IELTS listening subscore of 7.7 (*SD* = 0.8). Self-reported proficiency from scalar data (see Appendix: Language Background Questionnaire) showed L2 participants identified as advanced-18, high intermediate-19, intermediate-8, and low intermediate-2 for overall English ability. For listening, participants indicated their ability as advanced-19, high intermediate-17, intermediate-8, and low intermediate-2. Mean age of arrival was 24.3 years (*SD* = 5.7). While not directly recruited from an English teaching cohort, 25 participants reported having taught English for an average 2.4 years (*SD* = 3.4); 21 reported no English teaching experience.

Mandarin was selected for its empirically documented and theoretically explained tendencies to conflate target vowel pairs (see Appendix: Language Background Questionnaire). The Spanish L1 group replaced the Korean L1 employed in Experiment 1. The adjustment was made to facilitate recruitment. Given the Spanish L1’s active presence in university societies, it was thought that there would be a larger pool of untapped participants to recruit from. The Spanish participants came from 5 different countries: Chile-8, Spain-3, Cuba-1, Mexico-1, Paraguay-1. Including Spanish L1 participants from several countries (Castilian and Latin American varieties) for vowel perception research corresponds with the literature (Flege et al., 1997, Flege & Wayland, 2019; Iverson & Evans, 2009). The Spanish L1 varieties spoken by participants in this study share a standard vowel inventory (/i/, /e/, /a/, /o/, /u/) with non-contrastive duration (Canfield, 1981; Ronquest, 2018).

Similar to Mandarin (and Korean) L1, the literature shows that Spanish L1 English speakers tend to conflate /i/ and /ɪ/ (Flege et al., 1991; Iverson & Evans, 2009), with Spanish L1 speakers assimilating the English phones into a single category, but unlike Mandarin, Spanish L1 have been shown to differentiate /ɛ/ and /æ/ (Flege et al., 1997). Consequently, we would expect Spanish performance to be similar to Mandarin speakers for the high vowel pair /i, ɪ/, and better for /ɛ, æ/.

Instrumentation

Language Background Questionnaire. The language background questionnaire described in Experiment 1 Methods was used for Experiment 2, but reformatted for online administration.

Listening materials. Listening prompts consisting of isolated words and connected speech were developed to assess listeners' ability to differentiate between target vowels (/i, ɪ/ and /ɛ, æ/). Materials consisted primarily of a subset of prompts developed for, and explained in, Experiment 1. A reduction of listening materials was required to offset the expected physical and durational demands of transcription. Isolated words beyond bVt (i.e., Diverse Words) were excluded and for sentence prompts, only two of the four talkers were used with each pair (e.g., Talker 1 and 2 might be used for Sentences 11 and 12, while Talker 3 and 4 might be used for Sentences 13 and 14). Prompts included isolated and connected speech. Isolated speech (bVt) was used in oddity and transcription tasks; connected speech was only used in transcription tasks and included Travel Agent, Question and Answer, and Diverse Sentences. The Experiment was built and administered using Gorilla.sc, the platform described in Experiment 1 Methods.

Graded complexity of listening prompts. The blocks of listening prompts were designed to incorporate variability with graded complexity. Tasks included discriminating between bVt words in isolation, identifying bVt words in isolation, listening for specific information in a predictable sentence pattern (Travel Agent task) and in a non-predictable sentence pattern (Question and Answer task), and finally to listening to an entire utterance, where no focus was provided regarding target information, and participants typed the sentence they heard on the computer keyboard (Diverse Sentences). Table 11 summarises prompt complexity.

Table 11. Graded complexity of prompt types for Experiment 2

	Prompt type				
Feature	Oddity	bVt identification	Travel Agent	Question & Answer	Diverse Sentences
Tokens[1]	multiple (2-3)	One	One	one	one
Target word	bVt	bVt	diverse words	diverse words	diverse words
Focused	yes[2]	Yes	yes (question directed)	yes (question directed)	no
Syntax	predictable, isolated words	predictable, isolated word	predictable, target word in sentence final position	unpredictable, target word in any position	unpredictable, target word in any position

Note. Complexity is shown from least (leftmost prompt type column) to greatest (rightmost column).

[1] Tokens indicates number of target words (tokens) per prompt. Oddity contained either 2 or 3 tokens of a target vowel: 2 where an “odd” word was present, 3 where all tokens were “same”.

[2] The oddity task was considered focused as each word in the sequence of three was spoken in isolation (i.e., not in a sentence) and participants were able to listen for specific differences in vowel characteristics, such as quality or duration.

Development considerations

Memory, spelling, and replacement. Additional care had to be taken in selecting sentences for Experiment 2 because transcription tasks contend with potential confounds such as working memory, spelling, and replacement (intruder words). Findings from the literature helped guide the refinement of the sentence lists to reduce memory load and facilitate short-term recall. First, length was considered; too long of a sentence would lead to floor effects (Armon-Lotem et al., 2015) if participants could not recall sufficient information to transcribe the target word (which was not always known as the target word to participants). Bley-Vroman and Chaudron (1994) quantified the inverse relationship between sentence length and recall, finding an increasing decline from seven words onward (ceiling effects remained up to six words). Additionally, L2 learners may more readily attend to content words than function words (Field, 2008), making the number of content words salient. With this knowledge, the mean number of words in sentences used for Experiment 2 was 5.3 (*SD* = 1.2) for /i, ɪ/ and 5.0 (*SD* = 1.7) for /ɛ, æ/. Content words were restricted to a maximum of five, with a mean of 2.9 content words for both /i, ɪ/ (*SD* = 0.8), and /ɛ, æ/ (*SD* = 1.1).

Elicitation tasks—utilising sentence repetition—have been used to predict proficiency (Gaillard & Tremblay, 2016), with best performing participants on elicitation tasks correlated with the highest proficiency levels (as determined by cloze test scores) and the lowest performing participants correlated with the lowest proficiency. Combining the high intermediate to advanced proficiency of the current sample with the restricted count of total and content words in Diverse Sentences, participants were expected to be comfortably within range for felicitous recall.

Paraphrasing or “conceptual representations” were threats to verbatim reproductions (Potter and Lombardi, 1990), as participants may replace words with synonyms as they regenerate sentences from short-term memory. Though a paraphrased response would be evidence of accurate perception, in the instance where a participant did not properly hear or know the target word, responses may be uninterpretable. For instance, a participant may have accurately perceived the target vowel and word, but the replacement word might not be one I, as the researcher and grader, might associate with the word or context; likewise, the participant may have misperceived the vowel and word, but the synonym used might be considered correct due to a figurative interpretation or benefit of the doubt. Therefore participants were instructed to transcribe words verbatim. Supporting and extending the concern of substitutions, Antonijevic et al. (2017) found that for sequential bilinguals, the L2 may be more likely prone to omission errors than substitution errors (compared to the L1 which is more prone to substitution), though both were present in sentence repetition tasks. A decision to monitor both substitutions and omissions was made, and these errors were included in a transcription coding scheme.

Spelling was another anticipated confound. Prior to the experiment, participants were informed that they may encounter a word they did not know or did not know how to spell. Participants were instructed to spell such words exactly as they sounded, both in the instructions and after practice sessions of each sentence transcription task (block of listening prompts). L2 participants were given the benefit of the doubt for spelling, resulting in ambiguous interpretations (e.g., not following a double consonant rule, thus spelling a different word with a different vowel sound); L1 participants were not. (See Transcription Coding Scheme for further explanation.) Due to this guideline for extending benefit of the doubt for the L2 group, it was possible that it inflated the final L2 group scores and that the gap between the Control and the target L2 groups was slightly[44] larger than reported. To reduce the need for this benefit of the doubt, where possible, I clarified participants' ambiguous responses in a post-experiment debrief.

Suitability of the sentence prompts was demonstrated through a preliminary administration among peers and a two-person Experiment 2 pilot; however, as these results were cursory, the ability for participants to recall information faithfully and relate target words was monitored throughout the experiment.

[44] Results across all transcriptions showed that there were relatively few instances where a benefit of the doubt had to be given.

Word frequency, familiarity, and association. Controlling for lexical frequency was initially a consideration as participants may conceivably be prone to write words they were most familiar with, thus creating a response bias. If a participant who had difficulty perceiving the difference between /ɛ/ and /æ/ heard, “It will be an expansive study” but did not know the term “expansive”, only “expensive”, they may predictably type, “It will be an expensive study”. If the participant later heard, “It will be an expensive study”, they would also be expected to type “expensive”, and during analysis it would be impossible to uncover whether a correct response was due to accurate perception or merely a function of familiarity.

Word frequency, however, was not previously found to be a reliable correlate for perception (see Results, Experiment 1). Two factors may explain this to help improve the measure for Experiment 2: lack of localisation to a specific group or individual and the inability to account for context. Frequency lists from corpora typically offer general measures of prevalence in L1 contexts and therefore may not sufficiently reflect a target L2 group. Employing a frequency list designed for a specific L2, such as one extracted from a learner corpus, may give a broad idea of an average L2 learner, but still cannot guarantee accurate characterisation of the experiment’s particular sample or learner. Word knowledge, indicated previously, is important, but it would have to be localised to individuals within the sample. Given the divergent countries, educations, and experiences of the experiment’s participants, participants’ individual familiarity with words would be required to obtain a meaningful connection between vocabulary and perception. Eliciting participants’ individual vocabulary knowledge as it relates to target words would resolve the lack of localisation encountered using frequency lists.

Familiarity alone, however, may still not adequately account for the tendency to hear one word over another as familiarity does not entail association with a given context. For example, in the carrier sentence, “calculate the [betting/batting] averages”, someone who is familiar with baseball may be expected to associate “batting” with the provided sentential context, while someone who is unfamiliar with baseball may more strongly associate “betting” with the context. Consequently, determining association of a word with the provided sentential context may provide additional, if not more revealing insight into the interplay between lexical knowledge and perception.

Word lists.

bVt. The isolated bVt words, beat, bit, bet, and bat, were used in discrimination and identification tasks as a comparison for listener performance with sentential listening prompts. The bVt frame was chosen for its historically and current prominence in the literature (see A brief history of vowel perception assessment).

Travel Agent. The Travel Agent sentence list was created to replace the Directions sentence list from Experiment 1. The rationale for the change was to develop balanced number of items for the vowel pairs. The study included real words and real-world locations to create minimally paired listening prompts, and with Directions (a street name word list) fewer locations were obtained for the /i, ɪ/ vowel pair than /ɛ, æ/. It was found that a more balanced list could be developed by using UK regions rather than London streets. As with Directions sentences, the Travel Agent sentences began with a fixed carrier phrase followed by a location. The carrier sentence was "Book us a room in [location name]". Six new minimally paired sentences were developed and recorded by the same Talkers and in the same manner previously described. The final Travel Agent list comprised of four minimally paired sentences for each vowel pair, displayed in Table 12.

Table 12. UK-based locations used in sentences for Travel Agent task

Target vowels	Cognate Location	Cognate Location
/i, ɪ/	Leece	Liss
	Leverton	Liverton
	Wheatfield	Whitfield
	Wheelton	Wilton
/ɛ, æ/	Brendon	Brandon
	Ecton	Acton
	Ester	Aster
	Henbury	Hanbury

Note. Location names completed the carrier sentence, "Book us a room in…".

Question and Answer. Sentences for the Question and Answer task (Question and Answer) were paired with questions which directed listener focus toward target words. One of the sentences for each vowel was also present in the Travel Agent word list. Sentences, target words, and questions are compiled in Table 13.

Table 13. Questions and listening prompts used in the Question and Answer task

Target vowels	Guiding question	Sentence
/i, ɪ/	Where does the speaker want to meet you?	Meet me at Radcliffe/Redcliffe Square.
	What averages should you calculate?	Calculate the batting/betting averages.
	What should you look for?	Look for chromosome banding/bending.
	The speaker wants to find a shop that sells what?	I'd like to find a shop that sells jams/gems.
	What did the athlete do?	The athlete lapped/leapt everyone.
	Where does the speaker want to book a room?	Book us a room in Brandon/Brendon.
	What did the speaker say?	I said bat/bet.
/ɛ, æ/	Where does the speaker want to meet you?	Meet me at Siemens/Simmons Road.
	What should you do for two laps?	Heat/hit the pool for two laps.
	What did the old man do?	The old man wheezed/whizzed past me.
	What does the elderly man not want to do?	The elderly man doesn't want to leave/live.
	What does the speaker say about impoverished people becoming leaders?	Many impoverished people have reason/risen to become leaders.
	Where does the speaker want to book a room?	Book us a room in Wheatfield/Whitfield.
	What did the speaker say?	I said beat/bit.

Note. Target words are underlined.

Diverse Sentences. Sentences were reduced from the larger set used in Experiment 1. As sentence length has been shown to have an effect on perception (Holt & Wade, 2004), sentences were limited to five content words (*M* = 3.0, *SD* = 0.9) and 9 total words (*M* = 5.1, *SD* = 1.5). Number of syllables was also considered (Spoehr & Smith, 1973), both for syllables in target words (*M* = 1.6), *SD* = 0.7) and the global sentence (*M* = 6.7, *SD* = 2.0). Each vowel pair contained 12 minimally contrastive sentences, totalling 24 unique sentences per pair. A summary of the sentences and their word and syllable count is provided in Table 14.

Table 14. Breakdown of sentences, syllables, and word count for Experiment 2 Diverse Sentences

Target vowels	Sentences and parts of speech[45]	Content words[46]	Total words	Target word syllables	Sentence syllables
/i, ɪ/	I said beat/bit. PRN V V	2	3	1	3
	The man was beaten/bitten. DET N AUX V	2	4	2	5
	Feel/fill the cavity first. V DET N ADV	3	4	1	5
	Take the lead/lid for me. V DET N PREP PRN	2	5	1	5
	The elderly man doesn't want to leave/live. DET ADJ N AUX NEG V PREP V	4	7	1	10
	The Dutch have basic meals/mills. DET N V ADJ N	3	5	1	6

Note. Target words are underlined with parts of speech labelled below each sentence.

[45] Labelling parts of speech enables a classification of functional and content words.

[46] To identify (and therefore count) which words were content words, it was necessary to classify which words were functional words. Function words are a relatively small, "closed class" of words (i.e., they do not change). However, despite consisting of relatively few, unchanging words, there is no consensus as to which precises words constitute function words, leading to a range of claims such as the list constituting "about 450" words (Hanon, 2015) and "a working figure, including many rare terms, is 300" (Weber, 2006). For clarity, this study characterised function words as determiners (articles, possessive pronouns, demonstratives), prepositions, pronouns, conjunctions, and auxiliary verbs. All other words were classified as content words.

	A peel/pill will clear your skin up. DET N AUX V DET N PREP	3	7	1	7
	Scheming/skimming helps in this profession. N V PREP DET N	3	5	2	8
	Meet me at Sheep/Ship Lane. V PRN PREP N	3	5	1	5
	Patients keep sleeping/slipping on the floors. N V V PREP DET N	4	6	2	8
	Book us a room in Wheatfield/Whitfield. N PRN DET N PREP N	3	6	2	7
	The old man wheezed/whizzed past me. DET ADJ N V PREP PRN	3	6	1	6

Target vowels	Sentences and parts of speech	Content words	Total words	Target word syllables	Sentence syllables
/ɛ, æ/	Meet me at Allen/Ellen Street. V PRN PREP N	3	5	2	6
	Locate the axons/exons. V DET N	2	3	2	5
	I said bat/bet. PRN V N	2	3	1	3
	Calculate the batting/betting averages. V DET ADJ N	3	4	2	9
	Book us a room in Brandon/Brendon. V PRN DET N PREP N	3	6	2	7
	I said cattle/kettle. PRN V N	2	3	2	4
	Locate the afferent/efferent neuron. V DET ADJ N	3	4	3	8
	It will be an expansive/expensive study. PRN AUX AUX DET ADJ N	2	6	3	9
	I'd like to find a shop that sells jams/gems. PRN AUX V PREP V DET N PRN V N	5	9	1	9
	Take a biopsy of that mass/mess. V DET N PREP DET N	3	6	1	8
	Those kayaks come with paddles/pedals. DET N V PREP N	3	5	2	7
	Critics panned/penned several recent articles. N V ADJ ADJ N	5	5	1	10[1]

Note. ADJ = adjective; ADV = adverb; AUX = auxiliary verb; DET = determiner; N = noun; NEG = negative; PREP = preposition; PRN = pronoun; V = verb.

[1] "several" was spoken as 2 syllables opposed to 3.

Listening Tasks

Discrimination (bVt Oddity). A single discrimination task, 3-interval oddity, incorporated the same audio as Experiment 1 (see Experiment 1 Methods). For oddity trials, three words were presented, with one word minimally phonetically contrastive with the other two (e.g., bet-bat-bet). This task diverged from Experiment 1 in having one less interval (word) for participants to listen to and having an added "same" option, where all words were lexically identical (e.g., bet-bet-bet) despite being physically different (spoken by different talkers). Eliminating the fourth token in oddity tasks resulted in a 25% reduction in playing time for each item in this block of prompts. Twenty-five percent of all oddity prompts were sequences of three lexically identical words. Replacing one of the word tokens with an option for same responses permitted a reduction of word tokens in each trial while maintaining a 25% rate for correct chance responses. Further, the "same" trials (catch trials) tested how well listeners were able to disregard non-contrastive, within category differences in target vowels while attuning to contrastive differences between vowels and permitted false alarm data to be computed (Best et al., 1981). Word sequences were balanced for Talker and place of correct answer, meaning the correct answer was Button 1, Button 2, Button 3, and Button 4 (where Button 4 was always "same") an equal number (*n* = 8) of times. Items were randomly administered. Four practice items were provided to acquaint users with the task, with correct answers elicited for each of the four on-screen buttons.

Identification (transcription). Experiment 2 complements Experiment 1's identification tasks by employing an open response type (i.e., the option was not explicitly labelled and participants are free to write whichever word they believe was played in the prompt)—orthographic transcription. Having an open response type prevented the possibility for participants to ignore sentences to focus on specific words and greatly mitigated the potential for correct chance responses. English orthography, opposed to a phonological system such as the International Phonetic Alphabet, was used to avoid additional training for participants. For the present study, accuracy focused strictly on target vowels, ignoring spelling to the extent possible (see Transcription preparation), and using the number of correctly identified vowels as the listeners' perceptual score. This is similar to the "word match technique" used in intelligibility research (Derwing & Munro, 1997; Derwing et al., 2002), where intelligibility is determined by the proportion of correct words identified, ignoring trivial errors. Spelling, so long as it was readily identifiable as the intended word, was trivial for Experiment 2. Where a target vowel was not readily identifiable by the transcription, this was considered a non-trivial error. Like Derwing and Munro (1997), both trivial and non-trivial errors were coded, but only non-trivial errors counted against the listeners' perceptual accuracy scores.

Item counts. Item count was truncated to ensure participants could complete the experiment in a single sitting in approximately 90 minutes or less. Table 15 summarises item count by prompt type and vowel pair.

Table 15. Experiment 2 item count by type and vowel pair

Task Block	/i, ɪ/ items	/ɛ, æ/ items
bVt Oddity	32	32
bVt transcription	8	8
Travel Agent	16	16
Question and Answer	28	28
Diverse Sentences	48	48

Practice. Practice listening prompts were included for each task, serving two important roles. First, it familiarised participants with the tasks and ensured they were cognisant of what was expected. Second, it helped calibrate listeners to each talker's voice, promoting talker familiarity and normalisation. Previous studies have shown talker variability can have a negative impact on speech recognition and recall (Nusbaum & Morin, 1992). Due to the short-term memory requirements of the current study, where participants must accurately recall entire utterances, this directly impacts participants. Practice items, it was reasoned, would help counterbalance this deleterious effect on memory as perceptual adaptation can occur within a few sentences (Clarke & Garrett, 2004). Further, familiarising participants with talkers' acoustic characteristics provides auditory cues which may help them resolve ambiguous phonemes (Uddin et al., 2020). Doing so in advance, at least in part, will promote reliability of results as adjustments are made prior to formal assessment. Four practice items were provided per block[47].

Associations (self-report survey). The association task investigated which word in a minimal pair participants associated with each sentence. Sentences were textually displayed on screen, one at a time, with an underline replacing a target word. A slider bar was presented beneath the sentence and words from the minimal pair were at polar ends of the slider. Sounds were consistently placed so that words with /i, ɛ/ vowels were oriented on the left end of the slider and /ɪ, æ/ on the right. Participants were asked which word they associated with the sentence context and instructed to use the slider accordingly. Participants were informed that there was no inherent correct answer; the only correct answer was the one which reflected their personal associations of the words with the sentences. If no word was associated over another for the context, participants were able to leave the slider at the default neutral position, equidistant from the two words. This task took approximately four minutes to complete.

[47] Derwing and Munro (2002) offered two “warm-up items” before tasks to provide “a sense of the range of accentedness”, suggesting relatively few items may be necessary to acquaint participants with accent.

Word familiarity (self-report survey). A short survey (approximately 3.5 minutes) was administered to listeners to investigate the potential interaction of vocabulary with perceptual performance. The survey was adapted from Paribakht and Wesche's (1997) work on word knowledge. Paribakht and Wesche's model explored word knowledge through self-report data, asking participants to write synonyms or the meaning of words, and then analysing interpretations. To minimise participant effort and automate analysis, the adapted version employed in this experiment used a numeric scale to reflect the extent of knowledge. The ordinal scale consisted of the following five descriptors:

1. I don't remember hearing this word before.
2. I have heard this word before, but I don't know what it means.
3. I have heard this word before and I think I know what it means.
4. I have heard this word before and I know what it means, but I can't use it in a sentence.
5. I have heard this word before and I know what it means, and I can use it in a sentence.

Reducing the time required to complete the experiment, vocabulary terms related to proper nouns (e.g., Allen Street, Ellen Street) were excluded from the vocabulary self-report. It was reasoned that sufficient coverage of terms was present to conduct the desired analyses without including proper nouns. For Diverse Sentences, the primary focus of connected speech analyses, this left 40 of 48 vocabulary terms.

Participant Experience.

A central focus of the current research is exploring the suitability of sentential tasks for the purpose of instrument development. In-line with best practices in instrument design (Backman & Palmer, 1996, 2010), suitability included information about participant experience from the participant perspective. Consequently, a survey was appended to the end of each block of listening prompts and an experiment-final question asked participants about their experience.

NASA Task Load Index (NASA TLX) questionnaire. The NASA TLX (Appendix: NASA TLX screenshot) was employed as a means of quantifying the effect of added complexity in prompt types from a participant's subjective perspective. This was valuable to explore as L2 participants are stakeholders in vowel perception assessments. The pilot questionnaire was undesirable as it was inadequately tested and refined. Further, it was intended for use at the end of the experiment, and therefore elements of the experience with preceding prompt types might no longer be salient. The NASA Task Load Index (TLX) addressed the limitations of the pilot questionnaire.

The NASA TLX is a widely cited instrument (Grier, 2015) designed to rapidly and unobtrusively obtain work load information during or immediately after a task when it was most salient, making it ideal for extended experiments with multiple tasks (Hart & Staveland, 1988). The NASA TLX consists of six self-reported scalar measures of mental, physical, and temporal demands of each task, along with effort, frustration, and estimated performance. The original NASA TLX contains two parts. In the first section, participants rate each subscale (e.g., mental effort). In the second section, there is a weighting component where participants must provide relative weighting for each subscale. Congruent with the survey's most common modifications, the weighting component was not included (Hart, 2006). This made the questionnaire more intuitive for participants and faster to administer. The survey was further modified for uniformity, changing the labels 'excellent' and 'poor' to 'high' and 'low', matching the verbiage of other items. The modification altered the polarity of the item, a consideration for analysis. See Appendix Q for the revised survey.

The survey was appended to each block of listening materials. It included the primary question for each subscale (e.g., "How mentally demanding was the task?") prominently displayed above its label (e.g., "Mental Demand"). A slider bar was beneath the label, with the word "Low" to the left and "High" to the right. A draggable indicator was placed at the centre of the slider by default. As with the original NASA TLX, the slider contained 20 positions (steps).

Having the survey at the end of each task opposed to the end of the study ensured the task remains salient.

Open-ended question. Following Experiment 1 justifications and procedure, Experiment 2 offered an open-ended response prompt at the conclusion which encouraged participants to offer feedback regarding their experience. For Experiment 2, there were 29 included participant responses and a corpus of 2297 words. Code development and analysis are explained in Analysis and statistical approach.

Procedure

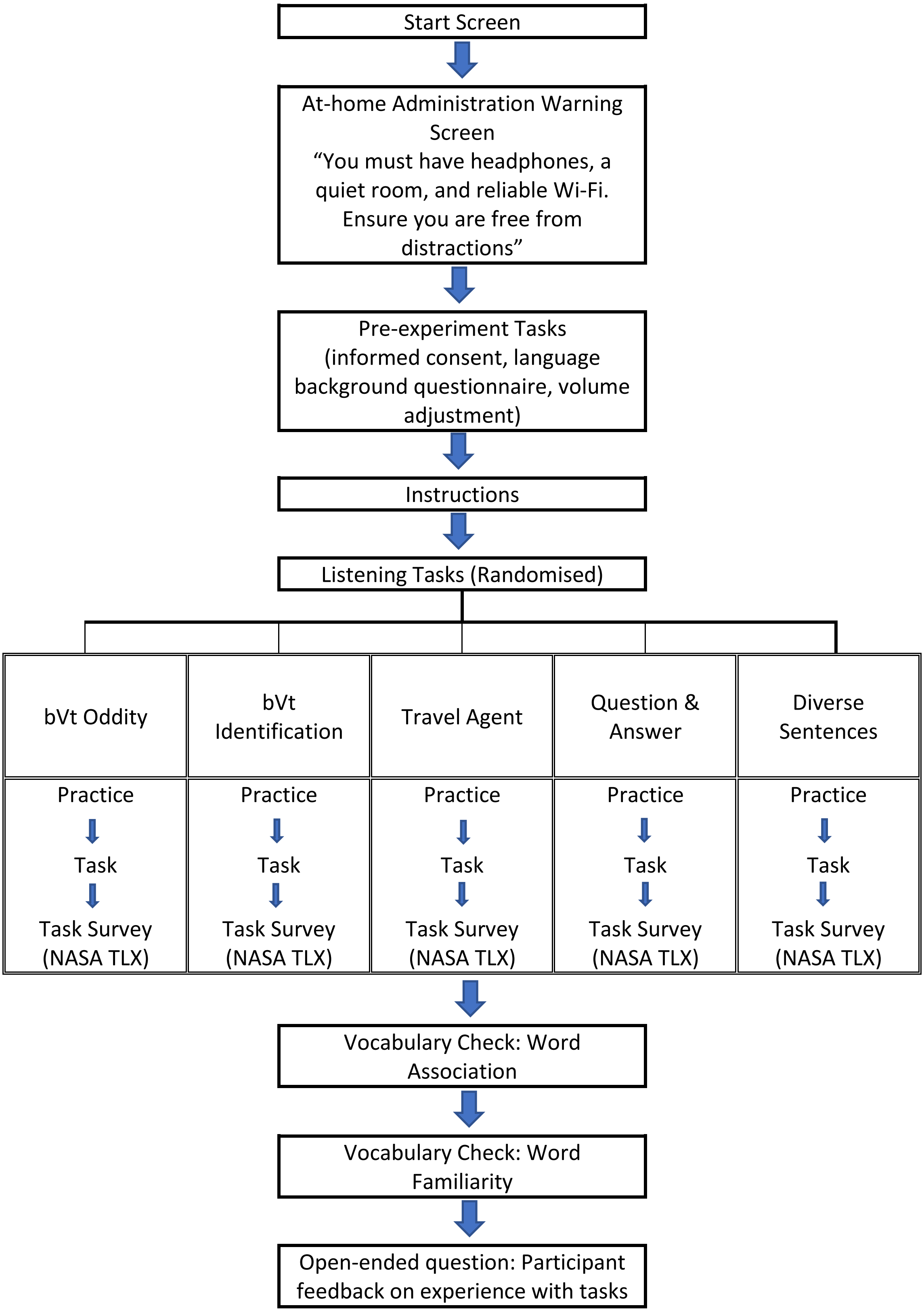
The study was conducted in one of two administration types, at-university and at-home (single session of approximately 90 minutes, including breaks). At-university administration was in a quiet room where I was present. At-home administrations were conducted at participants' homes from morning to early afternoon, local to participants. Times were scheduled and I was available online to respond to queries. Akin to Experiment 1, the experiment was administered using Gorilla. Immediately prior to the listening experiment, I explained the study's tasks and procedures, and that I would be present throughout the experiment. Participants completed an electronic consent form and language background questionnaire, and at-university administrations were given a pair of padded, on-ear headphones. Participants in at-home administrations used their personal headphones or earphones, but having a working pair (along with a quiet room and reliable WiFi) was a stated prerequisite for participation.

Before starting the listening experiment, participants were greeted with a volume adjustment screen which featured a generic, up-beat instrumental song. Once participants had confirmed their desired volume, they clicked next to begin the first of five blocks of listening prompts (items). Listening prompt blocks and the items within them were presented randomly and binaurally. Tasks were self-paced to enable audio replay (see Replay). A progress bar allowed participants to monitor their progress within that particular task and a message notified participants of the midpoint of each task. Upon completing each block of prompts, participants were prompted to complete a corresponding post-task survey.

After all tasks and their associated surveys had been completed, participants were given the opportunity to write additional notes about their experience in a textbox. For participant reference, the question listed each of the item types with examples.

Figure 5 provides a graphical summary of Experiment 2's data collection procedure.

Figure 5. Graphical outline of design and procedure for Experiment 2


Start Screen
At-home Administration Warning Screen
“You must have headphones, a quiet room, and reliable Wi-Fi. Ensure you are free from distractions”
Pre-experiment Tasks (informed consent, language background questionnaire, volume adjustment)
Instructions
Listening Tasks (Randomised)
bVt Oddity
bVt Identification
Travel Agent
Question & Answer
Diverse Sentences
Practice
Task
Task Survey (NASA TLX)
Practice
Task
Task Survey (NASA TLX)
Practice
Task
Task Survey (NASA TLX)
Practice
Task
Task Survey (NASA TLX)
Practice
Task
Task Survey (NASA TLX)
Vocabulary Check: Word Association
Vocabulary Check: Word Familiarity
Open-ended question: Participant feedback on experience with tasks

Analysis and statistical approach

This research investigates the effects of implementing diverse and sentential stimuli in testing vowel perception using an open-response prompt type (transcription). This section explains how transcription was prepared for analysis, outlines the research questions, and explains which analyses were used to answer each question.

Transcription preparation. Before analysis could be conducted, transcriptions had to be consistently and accurately marked. However, with the variability of transcriptions and the potential duration of the study, grading consistency was a concern. A coding scheme was thus developed to provide a transparent, reliable documentation of the results, which could readily be referenced throughout the study. The coding scheme is explained in Table 16.

Table 16. Transcription codes and explanations

Code		Explanation	Target word	Example Response
1	Right word	the vowel is correct and spelled correctly. Homophones and commonly accepted alternate spellings of a word are accepted as correct	Allan	Allen, Alan
2	Right word, wrong spelling	the spelling is off but the vowel and word are unambiguously correct	axons	acsons
3	Right vowel wrong word	the vowel is correct, the spelling is incorrect and the word is incorrect or unclear	jams	yams
4	Opposite target vowel in the minimal pair	the opposite vowel in the pair was reported	bat	bet
5	Wrong vowel wrong word	a vowel beyond one from the target pair was indicated, the word is wrong	mills	mails
6	Mixed or ambiguous	it is unclear whether the target vowels were indicated	efferent	French
7	No response, incomplete response	the response was uninterpretable due to lack of information	Meet me at Ship Lane	Meet me at…

While the codes employed numbers for convenience, they reflect value labels rather than numerics. Codes from 1-3 all denote a correct response and are worth a single point (i.e., 1), while 4-7 denote incorrect responses worth no points (i.e., 0).

An additional category, 0, was added to account for transcriptions which were ambiguous, but correct if given the benefit of the doubt. This code was motivated by the tendency for some participants to not add a double consonant at morpheme boundaries (e.g., bet-betting). Instead of interpreting “beting” as “beating” in the marking phase, it was interpreted as “betting”. The benefit of the doubt became more precarious when the incorrect spelling resulted in the formation of a new word, such as “bating” rather than “batting” or “Simon’s Road” instead of “Simmons Road”. Despite such instances, justification for giving the benefit of the doubt and maintaining the 0 code included (1) participants were instructed to spell words that they were unsure of as simply as possible, (2) English spelling rules may not come naturally to participants, particularly under uncertain conditions and in a research context, (3) the vowels in the new words were not generally expected to be conflated with the target vowel as they were typically spatially distant[48], diphthongal, or both (e.g., /ɪ/ to /ai/ in “whizzed” and “wized”), and (4) there were relatively few instances of these transcriptions. For instance, out of 5082 Diverse Sentences, 24 were labelled 0.

[48] Distant in vowel space—related to where they are formed in the mouth. Vowels with overlapping vowel spaces may be confused by L2 learners whose L1s do not have such distinction, while vowels which do not have overlapping vowel spaces are readily distinguished.

Q1. What are the measurable effects of employing diverse listening materials (phonologically diverse words in sentences) with an open response type for assessing English vowel perception in advanced L2 learners at a London university?

Q1a. Compared with bVt prompts, to what extent are open-response sentential prompts reliable? Complementing Experiment 1's investigation of reliability with closed-set prompt responses, Experiment 2 investigated the reliability of sentence prompts with an open-response type. An alpha of 0.7 and above was the target for internal consistency. As alpha is sensitive to the number of items in an assessment, an adjusted alpha measure is reported (in conjunction with the non-adjusted alpha), calculated based on an equal number of items across prompts. For the adjusted Cronbach's alpha, items were randomly selected for inclusion using the Microsoft Excel's RANDBETWEEN function. See Experiment 1 Methods for further explanation of the use and calculation of Cronbach's alpha for the present research.

Q1b. Compared with isolated bVt listening prompts, to what extent does prompt level-variability affect listener performance? Prompt level-variability refers to phonologically diverse target words in sentence-final position of a syntactically fixed carrier sentence (Travel Agent), question-focused listening for target words in syntactically varied sentences (Question and Answer), and sentence transcription where target words are not indicated (Diverse Sentences). The effect of variability was explored using proportion correct (*Pc*). Experiment 1 additionally used *d'*; however, due to the open-response type, *d'* was not suitable. *D'* explains decision making in identifying a binary signal (e.g., vowel 1 or vowel 2). The open response type meant that there was no second option to select or compare with the signal (vowel) which was presented.

Means were compared with a repeated measures ANOVA, and post hoc analysis were conducted for significant findings. The relationships between tasks were measured, exploring how well the Discrimination and Identification tasks correlate with each other, as well as how they correlate with performance in the primary connected speech task. Given previous vowel perception research, we may expect moderate-to-no correlation between the isolated speech tasks (Pisoni, 1973; Fry et al., 1962), but it remained unclear how well performance on the isolated speech tasks would translate to performance in connected speech tasks. It was possible one would outperform the other, or that together they would predict performance better than individually. Consequently, both hypotheses were considered.

Q1c. What is the effect of association (same, equal, opposite) on L2 performance with Diverse Sentence prompts? A primary aim of this second study was to uncover the effects of target word association when using sentences to investigate vowel perception. The effect of target word association in sentence prompts was explored descriptively and inferentially. Frequency counts for associations are identified, showing the number of participants who selected same, opposite, and equal for each language group. Mean performance (reported in Question 1) is revisited, categorising correct answers according to association. A preliminary regression scatterplot illustrates the relationship between association levels, isolated speech, and Diverse Sentences[49], leading to the next section (Q1d) which more deeply investigates the connection between association and Diverse Sentence performance through a generalised linear mixed model analysis.

[49] A correlational matrix was also performed, with results found in the Appendix.

Q1d. What is the effect of L2 performance with traditional vowel perception prompts (bVt identification and bVt discrimination) on performance with sentential prompts? How does it compare with the effect of association on performance with Diverse Sentence prompts? The predictive efficacy of bVt prompts and association were explored with a generalised linear mixed model design (GLMM). The GLMM development were constructed similarly to Experiment 1 with the lme4 package (Bates et al., 2015) for the R programme (R Core Team, 2021). Target vowel pairs were investigated separately. Obtaining a correct response was the dependent (outcome) variable, and a binomial link function was used to generalise the binary data to linear scale (Schäfer, 2017). Fixed effects (predictor variables) were bVt oddity and identification, familiarity, and association. Participant and items were included as random effects with random intercepts (Brekelmans et al., 2020). As with Experiment 1, model building was stepwise (Janssen, 2012), starting from a null model and incrementally adding predictors of interest. Models were compared using likelihood ratio tests and Akaike information criterion (AIC) as an estimate of prediction error (Verbyla, 2019). The optimizer tool, BOBYQA (Powell, 2009), was used to decrease convergence errors. Individual predictors were summarised and compared using odds ratios, where 1 indicated no relation between the predictor and the outcome, greater than 1 specified greater odds, and less than 1 indicated lower odds. Odds ratios were accompanied by 95% confidence intervals created with the R package, sjPlot (Lüdecke et al., 2021).

To use the data obtained from the Association survey in the GLMM, each item had to be systematically labelled for each participant to be used as a categorical predictor (i.e., a “factor”). This was done by first identifying which word a participant associated with the sentence (according to the Associations slider activity) and then comparing it with the target word (the “key”) which was present in the listening prompt that played. If the word that the participant associated with the sentence matched the key, it was marked, “same”. If the participant instead associated the cognate word of the minimal pair with the sentence (i.e., the participant did not associate the key word with the sentence context), then it was marked “opposite”. If the participant had no association of one word in the minimal pair over the other for the sentence context, it was marked, “equal”. In case 1, for example, a participant indicates that she associates “betting” (not “batting”) with, “Calculate the ___ averages”. In this fictitious example, Item 1[50] for the participant was, “Calculate the betting averages”. After the participant’s raw performance data has been collected, in a long format[51] data table, Item 1 will be labelled as “same” (under a column in the data table for association). In the same block of listening prompts for this participant, Item 2 was “Calculate the batting averages”. Item 2 for the participant will receive the label, “opposite”. In case 2, a different participant indicated no preference for betting or batting to fill in the blank, “Calculate the ___ averages”. For Item 1 and Item 2, his association would be labelled as “equal”. With the labels applied to each item for each participant, in R, the association variable was converted to a factor for the mixed model analysis.

Based on Experiment 1 odds ratios, a small relationship between bVt performance and Diverse Sentences was expected. While word frequency was not able to converge in Experiment 1’s GLMM analysis, the association data which is specific to participants’ own experience was expected to have a meaningful effect predicting performance with individual listening prompts in Diverse Sentences.

[50] Because items were presented in random order for each participant, item numbers used in analysis were arbitrarily assigned to listening prompts.

[51] Long format tables include more than one row for each participant, opposed to a wide format table which only permits one row per participant. Long format data is required for mixed models analysis.

Q1e. To what extent do known confounds affect listener performance? Known potential confounds for transcription and word familiarity were monitored. Frequency of transcription code occurrence and proportion of codes indicating uninterpretable observations—whether due to ambiguity, incompleteness, or an absence of response—are provided. Self-reported vocabulary knowledge (familiarity) is also reported. Frequencies for each of the five vocabulary knowledge descriptors are summarised by language group, and mode familiarity (1-5) of individual words within each minimal pair in Diverse Sentences is provided.

Q2. What are L2 participants' perceptions of using sentence prompts compared to isolated bVt prompts?

Q2a. What was the perceived task load (mental, physical[52], temporal, predicted performance, effort) of sentential prompts compared with bVt prompts? Perceived task load was investigated with a NASA TLX survey. Participants indicated their experience with each task (block of listening prompts) using a slider scale marked "high" on one end and "low" at the other. Values were not displayed for participants, but ranged between 0-20 depending on placement of the slider. By default, the slider's starting value was neutral (located at 10). Results were tabularly summarised and subscales from the index were correlated with performance using Spearman's rho.

Q2b. How does L2 subjective experience (as identified through an open-ended question) with sentence prompts compare with their experience with bVt prompts? An experiment final question asked participants to discuss salient elements of their experience with the experiment. Answers were analysed using a thematic analysis (Table 27).

[52] Physical demand was incurred in this Experiment as participants had to manually type responses for all blocks of identification prompts.

Code development. The coding development process for Experiment 2 differed from Experiment 1 in that it was a hybrid of inductive and deductive approaches (Fereday & Muir-Cochrane, 2006; Swain, 2018). The previous study provided a basic scaffold of coding; however, with the Diverse Sentences tasks, I expected that additional cognitions would be engaged. With the uncertainty of what the exact cognitions were and how they would impact scores and user experience, I viewed the data flexibly, allowing it to generate new codes or themes where appropriate. Coding revisions continued through training and calibration of the second coder.

I familiarised myself with the data by reading and rereading participant responses. Applying the basic code framework from Experiment 1, the original codes were a match, though I refined code definitions and added the theme, "Transcription".

Coder training. Coder 2 (discussed in Experiment 1), familiar with the research and coding, was again the external coder for Experiment 2. Training consisted of a PowerPoint presentation (see **Experiment 2 Training Presentation**) and calibration coding in NVivo. The presentation included motivations for Experiment 2 (a refresher on the research and the aims of the second study); revised vocabulary terms, explanations, and examples; the revised coding scheme; and a coding quiz. The quiz questions referenced all codes at least once and highlighted areas of potential disagreement (e.g., ensuring each code for cognition was paired with a prompt type, or not coding cognition without a prompt type present). Answers for the quiz questions and explanations for each were presented in the PowerPoint file. In lieu of an audio recording, I provided a narration script. Coder 2 played the presentation in PowerPoint's presenter view, enabling him to self-direct his progress.

I used excluded participant responses for calibration data, but modified the text to ensure all codes were represented. Coder 2 and myself completed the calibration exercise individually, but I was available to address Coder 2's questions before, during, and after the training activity.

Upon Coder 2's completion of the calibration exercise, I manually compared our coding, checking for agreement and disagreement. I then created a post-calibration Word document to summarise discrepancies (see Experiment 2 Appendix: Calibration notes). The document was clear to

Coder 2 and we were satisfied that differences were clarified and corrected. I sent Coder 2 the official NVivo file (a file with all response data and codes, but no reference links between the two) for the formal coding and we independently coded the data.

Code definitions and examples. Through an iterative process, described previously, I created the final coding scheme shown in Table 17.

Table 17. Experiment 2 qualitative codes, definitions, and examples used by coders

Code	Code definition	Data example (participants' verbatim comments)
Attention\focus and general attention	Reference to focus or attention which is not encompassed by fatigue, memory, or confusion.	*I found the bVt Transcription task a bit easier than the discrimination task because I only needed to focus on a single word.*
Attention\fatigue	Exhaustion, tiredness or weariness caused by specific prompts	*got a little bit tired when [doing] this part, but found it was easier to finish with context*
Attention\memory	Compromises to memory which impact participant attention.	*I couldn't recall words that I know…the sound of them didn't trigger my memory.*
Confusion	Uncertainty, particularly that which may misdirect or effect a participant's attention.	*If in the next sentence the speaker repeat the same [pronunciation], I might get confused".*
Emotion\negative affect	Negative emotion toward a prompt such as frustration, stress, anxiety, or aversion.	*I feel anxious about bVt Oddity, almost can't figure out the differences.*
Emotion\positive affect	Positive emotion toward a prompt such as approval or happiness.	*I loved the sentence transcription.*
General Cognition	Text which is relevant to cognition, but has no other code to describe it. Key examples include familiarity or miscellaneous thoughts about a prompt (or how the participant approached it).	*Sometimes I thought the sentence didn't make sense even I wrote down every single word (it seemed that the sentence did not follow a correct logic)*
Perceived difficulty	Any reference to difficulty, easy or difficult. This is perceived because participants may perceive a task as difficult but perform well, or perceive a task as easy but perform poorly.	*being a travel agent was a difficult task because of the names of the locations.*
Strategies	Methods participants use to answer the prompts beyond listening perception. Examples include using context to answer a question or guessing.	*I think this part is very subjective. Sometimes I guess the answers.*

Prompt\bVt oddity	Reference to the bVt Oddity task	*I found bVt Oddity easier to distinguish 'bet' and 'bat'*
Prompt\bVt transcription	Reference to bVt Transcription	*b-vowel-t transcription is a little bit difficult, but it is ok.*
Prompt\question and answer	Reference to Question and Answer	*The difficulty for me depended on the question - if it contains words I was not very familiar of then I would find it hard, otherwise it's ok*
Prompt\ Diverse Sentences	Reference to Diverse Sentences	*I found Diverse Sentences easy when I could understand what the sentence meant*
Prompt\travel agent	Reference to Travel Agent	*The unfamiliar place name to some extent impedes my answering.*
Transcription	All feedback text related to writing, typing, or spelling	*can not spell the words [in Travel Agent] with confidence*

Reliability. Reliability followed the same justifications and procedures as Experiment 1, incorporating Cohen's kappa coefficient and percentage of absolute correct to ensure consistency between coders.

Results

Preliminary quantitative analysis

The final sample included 33 Mandarin and 13 Spanish L1 listeners with 6 Control participants. Four participants were excluded from the original sample. Two speakers (1 Mandarin, 1 Control) had taken Experiment 1 and were excluded for potential testing effects, one Mandarin L1 participant was based in China rather than London (entailing an EFL rather than ESL demographic), and a final Spanish L1 was flagged for participant bias. The participant did not see the utility of bVt as a listening prompt[53] and consequently noted that she "didn't try" to respond accurately to these prompts (see Qualitative Results). Quantitative results for this participant were statistically irregular for bVt prompts and uninterpretable. This participant included qualitative data reflecting her opinions on the listening components, and this is discussed in Experiment 2 Qualitative Results.

Having both at-university and at-home administrations presented a complication to address before other analyses could be performed, as performance differences due to mode of performance

[53] explicitly stated both in a written response and post-experiment interview.

(in-person vs. online) could compromise the validity of the results. If differences in performance were found to be statistically non-significant, the two administration groups would be combined for analysis; if differences were found to be statistically significant, they would be analysed separately. An independent samples *t*-test measured the difference between at-home and at-university administrations of the experiment. The sample sizes for the two groups were unequal and thus Welch's t-test was performed (Delacre et al., 2017). The at-home group (n = 13, *M* = 68.2, *SD* = 11.5) was not significantly different (*p* > .05) than the at-university group (n = 33, *M* = 65.6, *SD* = 11.4). The lack of difference justified cumulative examination of the results.

Q1. What are the measurable effects of employing diverse listening materials (phonologically diverse words in sentences) with an open response type for assessing English vowel perception in advanced L2 learners at a London university?

Q1a. How reliable are sentential prompts compared to bVt?

Aligned with the purpose of exploring task functioning for practical instrument design, it was important to identify how internally consistent various tasks were. Table 18 provides a first look at each task (with all items), followed by an adjusted comparison in Table . The desired alpha level, as explained in Experiment 1 Methods, was set at .7 and above.

Table 18. Cronbach's α by item type (n = 46)

Vowel Pair	bVt Oddity (32 items)	bVt Identification (8 items)	Travel Agent (16 items)	Question & Answer (28 items)	Diverse Sentences (48 items)
/i, ɪ/	.89	-	.37	.76	.83
/ɛ, æ/	.85	-	.44	.83	.85

Cronbach's *α* is sensitive to the number of items in a scale, with larger item numbers generally relating to larger *α*. Consequently, *α* may be inflated for Diverse Sentences (48 items) compared with the two bVt tasks (32 and 8 items for Oddity and bVt Identification, respectively). For a more equitable comparison, item numbers were matched by randomly selecting 24 items each for bVt Oddity, Question & Answer, and Diverse Sentences. Randomisation was performed using the RANDBETWEEN function in Microsoft Excel, generating random labels of 1 or 0 for each item. Items marked with 1 were included while items with 0 were excluded. The bVt Identification and Travel Agent tasks had too few items and were not included in the 24-item comparison table.

Table 29. Adjusted comparison of Cronbach's *α* by item type (n = 46)

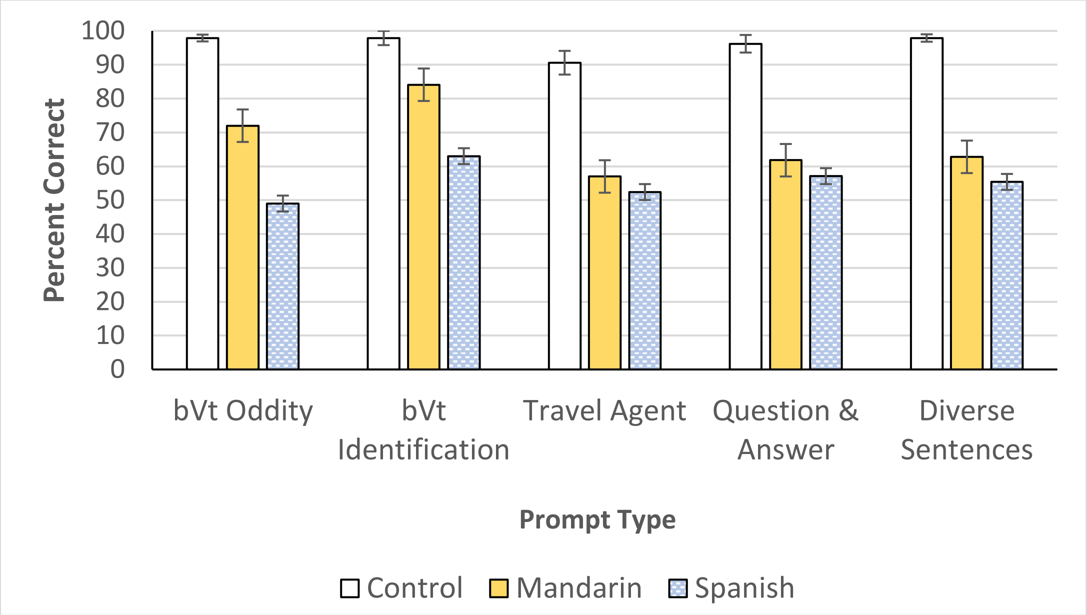
Vowel Pair	bVt Oddity (24 items)	Question & Answer (24 items)	Diverse Sentences (24 items)
/i, ɪ/	.86	.70	.75
/ɛ, æ/	.81	.78	.82

The adjusted 24-item results shown in Table suggest a favourable comparison of α across tasks and that fewer items could be employed while maintaining moderately strong internal consistency.

Q1b. Compared with isolated bVt listening prompts, to what extent does prompt level-variability affect listener performance?

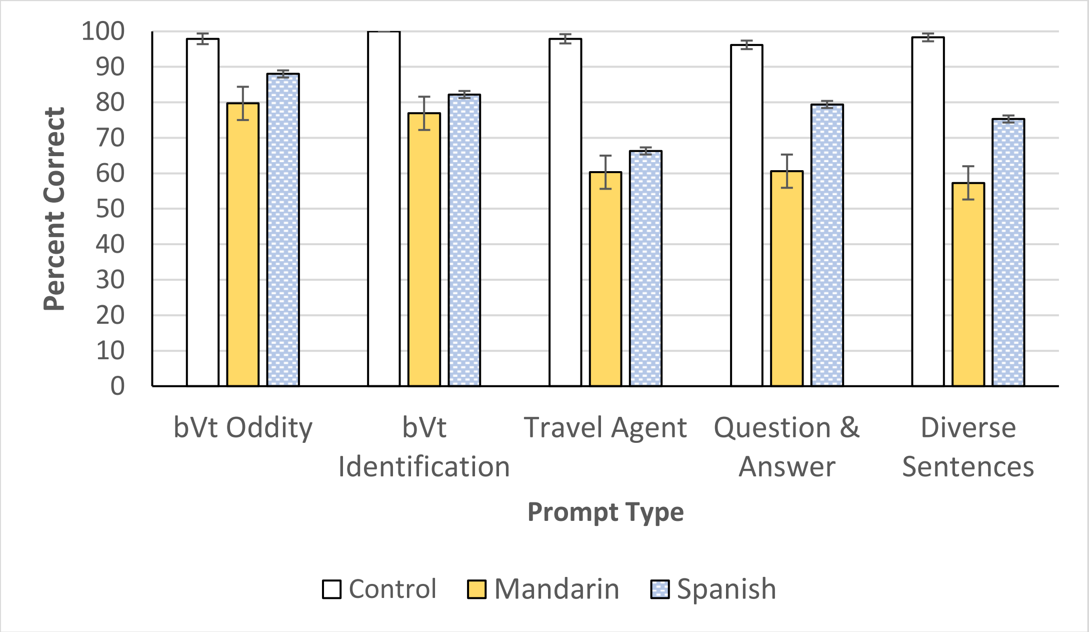
Mean performance functions as a descriptive indicator of whether the items and tasks performed as expected. The Control was expected to perform at or near ceiling levels for both vowel pairs while the Mandarin group was expected to perform nearer to 50%. The Spanish group was expected to perform closer to the Control for /ɛ, æ/, but closer to Mandarin for /i, ɪ/. Isolated word tasks (bVt Oddity and bVt Identification) were expected to be the easiest tasks according to percent correct, with Travel Agent, Question and Answer, and Diverse Sentences being increasingly more difficult. Results generally support these hypotheses, indicating efficacious task functioning.

Figure 6. /i, ɪ/ Mean Task Performance by Language Group



Note. Error bars display standard error. Chance performance for oddity was 25%. For identification, the response type was open and therefore chance remains unspecified.

Figure 7. /ɛ, æ/ Mean Task Performance by Language Group



Note. Error bars display standard error. Chance performance for oddity was 25%; For identification, the response type was open and therefore chance remains unspecified.

Differences between language groups were not examined with a multivariate analysis of variance due to the uneven sample sizes and relatively few Spanish and Control. Differences between performance on prompt type was, however, examined for the Mandarin L1 group.

Focusing strictly on the Mandarin L1, a repeated measures analysis of variance was conducted to determine whether the differences between tasks were statistically significant. Sphericity was violated for both vowel pairs and a Greenhouse-Geisser correction was consequently employed. Results were significant for both (/i, ɪ/ $F(2.8, 90.5) = 33.5$, $p < .001$, $\eta^2 = .51$) and /ɛ, æ/ ($F(2.5, 79.2) = 29.0$, $p < .001$, $\eta^2 = .48$). Post-hoc analysis showed a statistically significant difference between isolated speech tasks and connected speech tasks for both vowel pairs. For /i, ɪ/, differences were significant between all tasks while for /ɛ, æ/ the difference between isolated speech tasks was not found to be statistically significant.

Table 19. /i, ɪ/ Pairwise comparison of means by prompt type (n = 46)

Factors	Comparison Factors	Mean Difference	Standard Error	*p*
Oddity	bVt ID	-12.12*	3.34	<0.01
	Travel Agent	15.00*	3.22	<0.01
	Question & Answer	10.17*	3.39	0.01
	Diverse Sentences	9.21*	2.87	<0.01
bVt ID	Oddity	12.12*	3.34	<0.01
	Travel Agent	27.08*	2.29	<0.01
	Question & Answer	22.30*	2.40	<0.01
	Diverse Sentences	21.33*	2.38	<0.01
Travel Agent	Oddity	-14.96*	3.22	<0.01
	bVt ID	-27.08*	2.29	<0.01
	Question & Answer	-4.79*	1.99	0.02
	Diverse Sentences	-5.75*	1.95	0.01
Question & Answer	Oddity	-10.17*	3.39	0.01
	bVt ID	-22.30*	2.40	<0.01
	Travel Agent	4.79*	1.99	0.02
	Diverse Sentences	-0.96	1.75	0.59

Diverse Sentences	Oddity	-9.21*	2.87	<0.01
	bVt ID	-21.33*	2.38	<0.01
	Travel Agent	5.75*	1.95	0.01
	Question & Answer	0.96	1.75	0.59

Note. * indicates statistical significance.

Table 20. /ɛ, æ/ Pairwise comparison of means by prompt type (n = 46)

Factors	Comparison Factors	Mean Difference	Standard Error	*p*
Oddity	bVt ID	2.84	3.94	0.48
	Travel Agent	19.42*	2.27	<0.01
	Question & Answer	19.13*	2.84	<0.01
	Diverse Sentences	22.41*	2.61	<0.01
bVt ID	Oddity	-2.84	3.94	0.48
	Travel Agent	16.58*	3.24	<0.01
	Question & Answer	16.29*	3.12	<0.01
	Diverse Sentences	19.57*	3.36	<0.01
Travel Agent	Oddity	-19.42*	2.27	<0.01
	bVt ID	-16.58*	3.24	<0.01
	Question & Answer	-0.29	1.96	0.88
	Diverse Sentences	2.99	1.59	0.07
Question & Answer	Oddity	-19.13*	2.84	<0.01
	bVt ID	-16.29*	3.12	<0.01
	Travel Agent	0.29	1.96	0.88
	Diverse Sentences	3.28*	1.48	0.03
Diverse Sentences	Oddity	-22.41*	2.61	<0.01
	bVt ID	-19.57*	3.36	<0.01
	Travel Agent	-2.99	1.59	0.07
	Question & Answer	-3.28*	1.48	0.03

Note. * indicates statistical significance.

Results suggest that the two isolated speech tasks correlate as well or better with Diverse Sentences than with each other. Correlations between each task type are displayed as correlation matrices in Table 21 and Table 22.

Table 21. /i, ɪ/ Correlation matrix between prompt types (n = 46)

	bVt Oddity	bVt Identification	Travel Agent	Question & Answer	Diverse Sentences

bVt Oddity	1	.57**	.39**	.47**	.61**
bVt Identification	.57**	1	.45**	.51**	.56**
Travel Agent	.39**	.45**	1	.66**	.66**
Question & Answer	.47**	.51**	.66**	1	.83**
Diverse Sentences	.61**	.56**	.66**	.83**	1

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

Table 22. /ɛ, æ/ Correlation matrix (n = 46)

	bVt Oddity	bVt Identification	Travel Agent	Question & Answer	Diverse Sentences
bVt Oddity	1	.37**	.52**	.42**	.50**
bVt Identification	.37**	1	.50**	.46**	.43**
Travel Agent	.52**	.50**	1	.60**	.71**
Question & Answer	.42**	.46**	.60**	1	.88**
Diverse Sentences	.50**	.43**	.71**	.88**	1

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

Oddity correlated better with Diverse Sentences than with bVt Identification in both vowel pair correlation matrices. Such consistent findings may be indicative of Diverse Sentences concurrently tapping into distinct processes featured the two isolated speech tasks. This notion will be further explored in Discussion.

It was expected that Travel Agent would function as an intermediary task, sharing features with both connected and isolated speech, and would therefore correlate stronger with isolated speech prompts than Diverse Sentences. While supported with the low vowel pair, results were contra to expectations with the high vowel pair. Combined with aforementioned results, Travel Agent is a problematic task type despite its intended design to be the simplest connected speech prompt.

Cumulatively, the relationship between performance in isolated tasks and connected speech tasks is moderate, however the correlation is inefficient. While performing well with the isolated bVt words can indicate participants will also perform well in the more phonologically diverse connected speech contexts, the ease of the task meant that both strong and poor performers overall could score well on the isolated speech tasks. A stronger claim may be made that performing poorly with isolated speech prompts will translate to poor performance in connected speech. The use of association as a predictor of performance will be further explored in a generalised linear mixed model analysis.

Q1c. What is the effect of association (same, equal, opposite) on L2 performance with Diverse Sentence prompts?

This section shows raw frequency counts of each association level, connects association to mean performance, and investigates the relationship between associations, isolated speech tasks, and Diverse Sentences (further examined in Q1d's generalised linear mixed models).

Recall that "same" associations reflect congruence between the participant's association with the target word in its sentence context and the sentence which was played. In other words, whether the participant's association matches the key. "Opposite" indicates the word the participant associates with the sentence is not the target word, but the cognate word in the minimal pair. If neither word in the minimal pair is associated more strongly with the sentence that is played, the association is "equal". Not all participants indicated an association of "equal"[54]. Table 23 shows association frequency counts for /i, ɪ/ and /ɛ, æ/.

Table 23. Frequency counts for associations by group

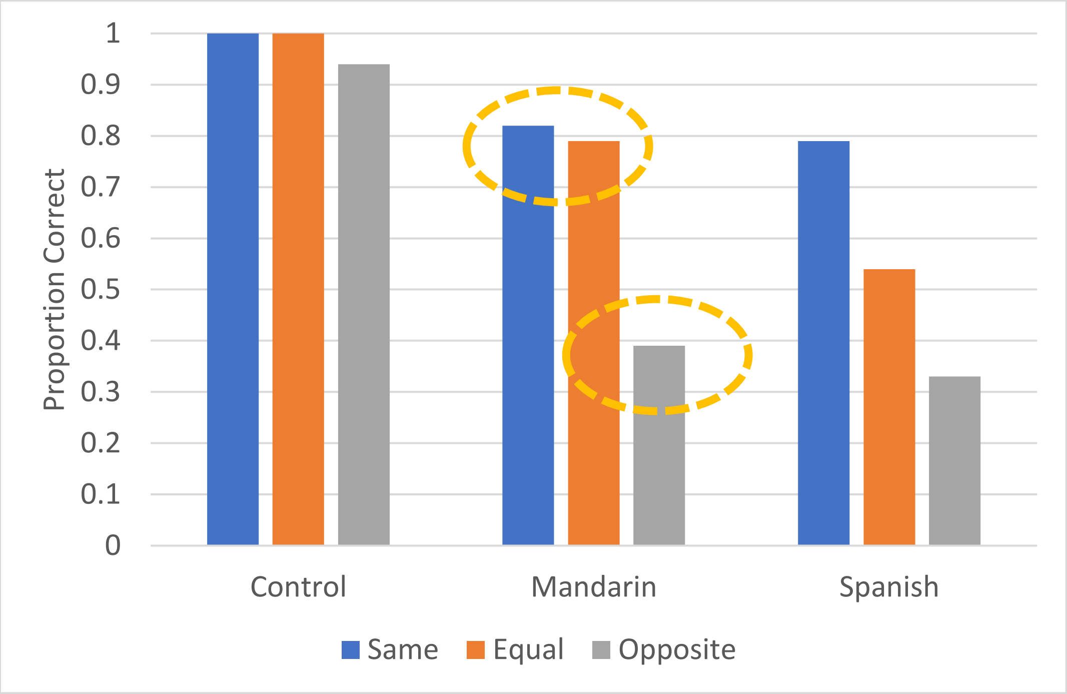
Vowel Pair	Group	Same	Equal	Opposite
/i, ɪ/	Control	107 (6)	72 (4)	108 (6)
	Mandarin	706 (33)	172 (14)	706 (33)
	Spanish	252 (13)	120 (10)	252 (13)
/ɛ, æ/	Control	111 (6)	64 (4)	111 (6)
	Mandarin	725 (33)	136 (11)	723 (33)
	Spanish	252 (13)	120 (9)	252 (13)

Note. () = number of participants who indicated the association.

[54] As both minimally paired words were possible for each carrier sentence, it was posited that more proficient speakers might have greater propensity to indicate an equal association. A *t*-test contrasted the difference in performance between people who indicated equal association (n = 25) and those who did not (n = 21). No significant difference was found in overall performance (*p* > .05). Further, no difference was found for IELTS score or self-reported proficiency level (*p* > .05).

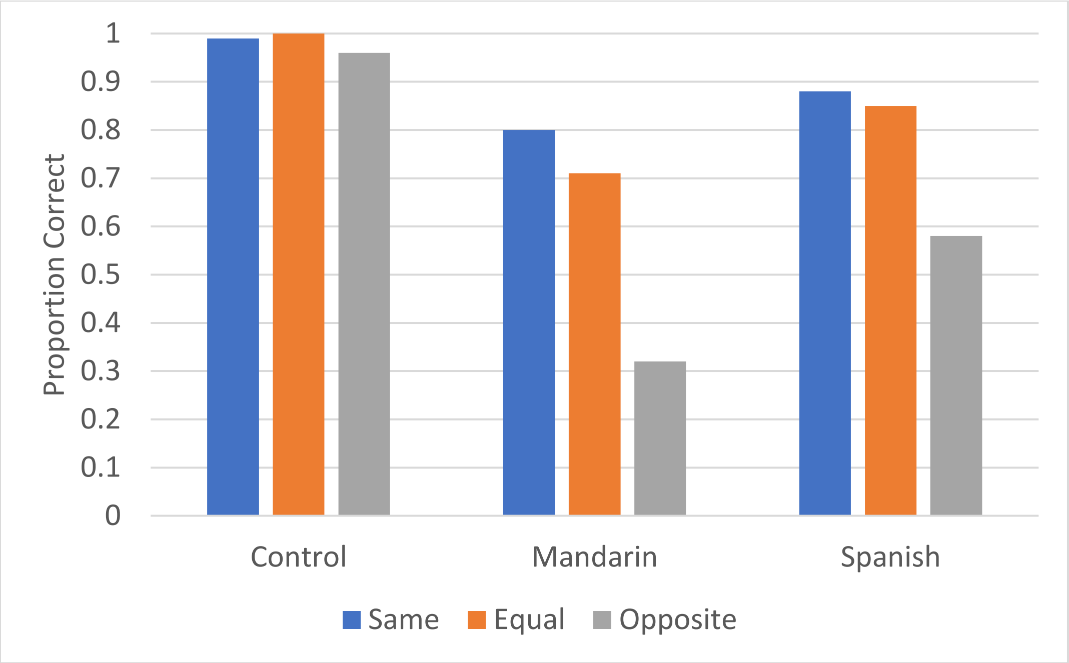
Revisiting mean performance. Mean performance provided a general overview of the data, but may be refined by examining performance through the lens of associations. Figure 8 and Figure 9 summarise the proportions of correct answers by association type.

Figure 8. /i, ɪ/ Cluster graph of proportion of correct answers by association for Diverse Sentences



Note. Associations: “same” reflects the participant associating the key (correct) word with the sentence it is heard in; “equal” means the participant associating both words in the minimal pair equally to the sentence context; “opposite” reflects the participant did not associate the key word with the sentence, but instead associated the other word in the minimal pair with the sentence.

Figure 9. /ɛ, æ/ Cluster graph of proportion of correct answers by association for Diverse Sentences



Note. Associations: "same" reflects the participant associating the key (correct) word with the sentence it is heard in; "equal" means the participant associating both words in the minimal pair equally to the sentence context; "opposite" reflects the participant did not associate the key word with the sentence, but instead associated the other word in the minimal pair with the sentence.

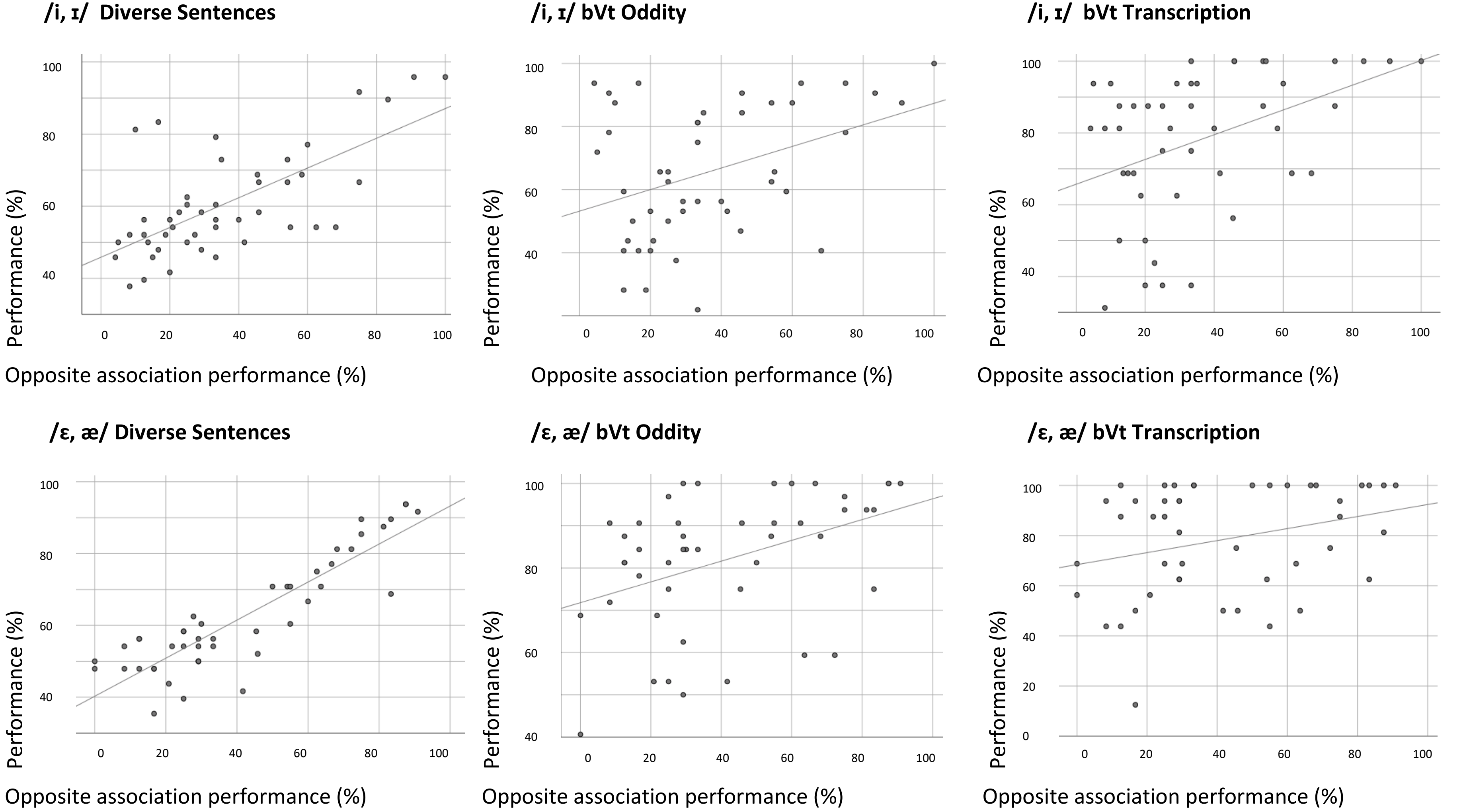
With association, there appears to be an interaction between top-down and bottom-up processing. Expectedly, the highest scores were generated from the "same" context, where participants associated the target word with the sentence in which it was heard. Correspondingly, it is also the least helpful measure as it is not possible to decipher whether the correct answer was due to accurate perception or association. The bottom-up process of vowel perception is confounded by the top-down process of context association.

The least perceptually biased control for sentence perception is "equal", where participants associate the minimally paired words equally with the carrier sentence. Performance with equal associations can be similar to same, which is proposed to be an inflated measure. Taking Mandarin as an example, the group performed reasonably well with neutral contexts for both the /i, ɪ/ (*M* =

79.1%, SE = 3.1) and /ɛ, æ/ (*M =* 70.6%, SE = 3.9). This appears to suggest effective bottom-up processing, but is limited and will be further explored in Discussion.

Preliminary associations. The relationship between opposite associations and the prompts bVt and Diverse Sentences was illustrated through a series of scatter plots in Figure 10.

Figure 10. Preliminary scatterplots of opposite association performance with Diverse Sentences by prompt type



The scatter plots exhibit that performing well on opposite association prompts for Diverse Sentences relates to performing well overall with Diverse Sentence prompts, and that ceiling effects occurred with the isolated speech prompts. With bVt prompts, results were mixed. At the lower levels of performance, associations were clearer as poor performance in isolated speech translates to performing poorly elsewhere. However, performing well in isolated speech did not necessarily lead to performing well in Diverse Sentences with opposite associations.

Table 24. High performers'[1] mean score (standard deviation) by prompt type for each association level on /i, ɪ/

Association	Oddity	bVt ID	Travel Agent	Question & Answer	Diverse Sentences
Opposite (n = 3)	92.7 (6.5)	100 (0)	77.1 (3.6)	92.9 (6.2)	93.8 (3.6)
Equal (n = 8)	92.6 (5.5)	95.3 (5.5)	75 (20.0)	87.5 (14.0)	90.6 (10.3)
Same (n = 27)	70.1 (21.1)	82.4 (19.1)	58.8 (11.0)	64.9 (15.5)	66.7 (15.2)

[1] High performance was determined by achieving at least 80% at the given association level. For instance, three participants scored 80% or above with opposite associations.

Table 25. High performers'[1] mean score (standard deviation) by prompt type for each association level on /ɛ, æ/

Association	Oddity	bVt ID	Travel Agent	Question & Answer	Diverse Sentences
Opposite (n = 6)	93.8 (9.7)	90.6 (15.7)	80.2 (12.1)	90.5 (10.5)	87.5 (9.5)
Equal (n = 9)	87.5 (16.5)	88.9 (17.3)	72.2 (18.0)	85.3 (13.8)	80.8 (11.7)
Same (n = 28)	85.3 (13.9)	83.5 (19.1)	66.2 (14.4)	72.2 (16.8)	68.7 (15.9)

[1] High performance was determined by achieving at least 80% at the given association level. For instance, six participants scored 80% or above with opposite associations.

Having established a preliminary distinction in performance between levels of associations, and having examined their relations to each other and isolated speech tasks, the next sections explores various associations and bVt prompts' propensity to predict perceptual performance in connected speech.

Q1d. **What is the effect of L2 performance with traditional vowel perception prompts (bVt identification and bVt discrimination) on performance with sentential prompts? How does it compare with the effect of association on performance with sentential prompts?**

This section combines previously expounded results to develop a working generalised linear mixed model which predicts connected speech performance. In Correlations, a moderate, positive relationship (*r* between .39-.61) was demonstrated between performance with isolated speech prompts and performance with sentential prompts. In associations, we uncovered a stark distinction in performance depending on association. Here the investigation is the relative strength of bVt prompts and associations to project responses on connected speech prompts.

To build a functional model, working, non-working, and superfluous variables were systematically identified. Initial models failed to run (converge). The complete model was saturated with fixed effects and failed to converge when both Familiarity and Associations were used as predictor variables. The model converged, however, when Familiarity was removed. This suggests the variance in Familiarity may already be explained by Association. Similarly, the model did not converge when Vowel Length was included. I reasoned that the random effect, Item, had already explained the variance provided with Vowel Length, thus "breaking" the model when R attempted to process the data. As a final omission, Language was removed. The model worked, yet the effect of Language was negligible and performance was contra to expectations. High vowel odds ratios were 0.04 for Mandarin and 0.05 for Spanish, extrapolating to slightly greater odds of Spanish participants responding correctly. This contradicted performance measures to that point as Mandarin had outperformed Spanish on every metric for the high vowel pair (both for raw and standardised scores). The irregularity was explained by the overlapping confidence intervals (0.01-0.12 for Mandarin; 0.01-0.18 for Spanish), meaning the difference was not statistically significant (Maltenfort et al., 2020). Given the diminutive odds ratios, the inconsistency of the finding, and that variance in

Language was likely largely explained by Participant, Language was removed as a predictor. This left a parsimonious model with three fixed effects (bVt Identification, Oddity, Association) and two random effects (Participant and Item).

Table 26 and Table 27 display results from the model building process, starting from the null model (m0) and progressively adding variables with random intercepts (m1-m4), and finally random slopes (m5-m6).

Table 26. /i, ɪ/ Model comparison for generalised linear mixed models

Model	Fixed effect	Deviance	*df*	AIC	LRT comparison	$X^2(df)$	*p*
m0	-	2584	3	2590			
m1	bVt Identification	2566	4	2574	m0-m1	17.2(1)	<.001
m2	Oddity	2551	4	2559	m0-m2	32.5(1)	<.001
m3	bVt Identification, Oddity	2546	5	2556	m2-m3	5.2(1)	<.001
m4	bVt Identification, Oddity, Association	2270	7	2284	m3-m4	255.7(2)	<.001
m5	**bVt Identification, Oddity, Association, Association random slope with Participant**	**2205**	**12**	**2229**	**m4-m5**	**64.2(5)**	**<.001**
m6	Association, Association random slope with Participant	2241	10	2261	m6-m5	35.2(2)	<.001

Note: Random effects included Participant and Item for all models. AIC = Akaike information criterion. LRT = Likelihood ratio test. Bold text reflects the best fit model.

Table 27. /ɛ, æ/ Model comparison for generalised linear mixed models

Model	Fixed effect	Deviance	*df*	AIC	LRT comparison	$X^2(df)$	*p*
m0	-	2615	3	2621			
m1	bVt Identification	2600	4	2608	m0-m1	15.3(1)	<.001
m2	Oddity	2597	4	2605	m0-m2	18.0(1)	<.001
m3	bVt Identification, Oddity	2591	5	2601	m2-m3	6.2(1)	<.001
m4	bVt Identification, Oddity, Association	2292	7	2306	m3-m4	299.0(2)	<.001
m5	**bVt Identification, Oddity, Association, Association random slope with Participant**	**2232**	**12**	**2256**	**m4-m5**	**60.6(5)**	**<.001**
m6	Association, Association random slope with Participant	2258	10	2278	m6-m5	26.0(2)	<.001

Note: Random effects included Participant and Item for all models. AIC = Akaike information criterion. LRT = Likelihood ratio test. Bold text reflects the best fit model.

Table and Table 28 exhibit results from the optimal models for the /i, ɪ/ and /ɛ, æ/ vowel pairs, respectively.

Table 39. /i, ɪ/ Model 5 (Optimal Model) Output

Predictors	Odds Ratio	Confidence Interval	*p*
Intercept	3.01	1.69 – 5.38	<0.001
bVt Transcription	1.03	1.01 – 1.05	0.011
Oddity	1.04	1.02 – 1.05	<0.001
Association—opposite	0.32	0.20 – 0.53	<0.001
Association—same	3.33	2.05 – 5.55	<0.001
N Participant	52		
N Item	48		
Observations	2495		
Marginal R2 / Conditional R2	0.29 / 0.60		

Note. Effect of bVt Transcription, Oddity, and Association on probability of participant transcribing the correct vowel during Diverse Sentence task

Table 28. /ɛ, æ/ Model 5 (Optimal Model) Output

Predictors	Odds Ratio	Confidence Interval	*p*
Intercept	3.78	2.14 – 6.68	<0.001
bVt Transcription	1.03	1.01 – 1.04	0.002
Oddity	1.03	1.01 – 1.05	0.006
Association—opposite	0.30	0.18 – 0.49	<0.001
Association—same	2.30	1.34 – 3.95	0.003
N Participant	52		
N Item	48		
Observations	2494		
Marginal R2 / Conditional R2	0.23 / 0.55		

Note. Effect of bVt Transcription, Oddity, and Association on probability of participant transcribing the correct vowel during Diverse Sentence task

The primary effects being investigated were the isolated speech tasks and association. Looking first at the isolated speech tasks as predictors of connected speech performance, we see a small (1.03), positive effect for isolated speech tasks on connected speech performance. The odds ratio of greater than one indicates that better performance on isolated speech prompts yields greater odds of a correct response with connected speech prompts[55]. The nearly identical odds ratios and confidence intervals for both isolated speech prompts suggest the tasks were equal predictors of connected speech performance.

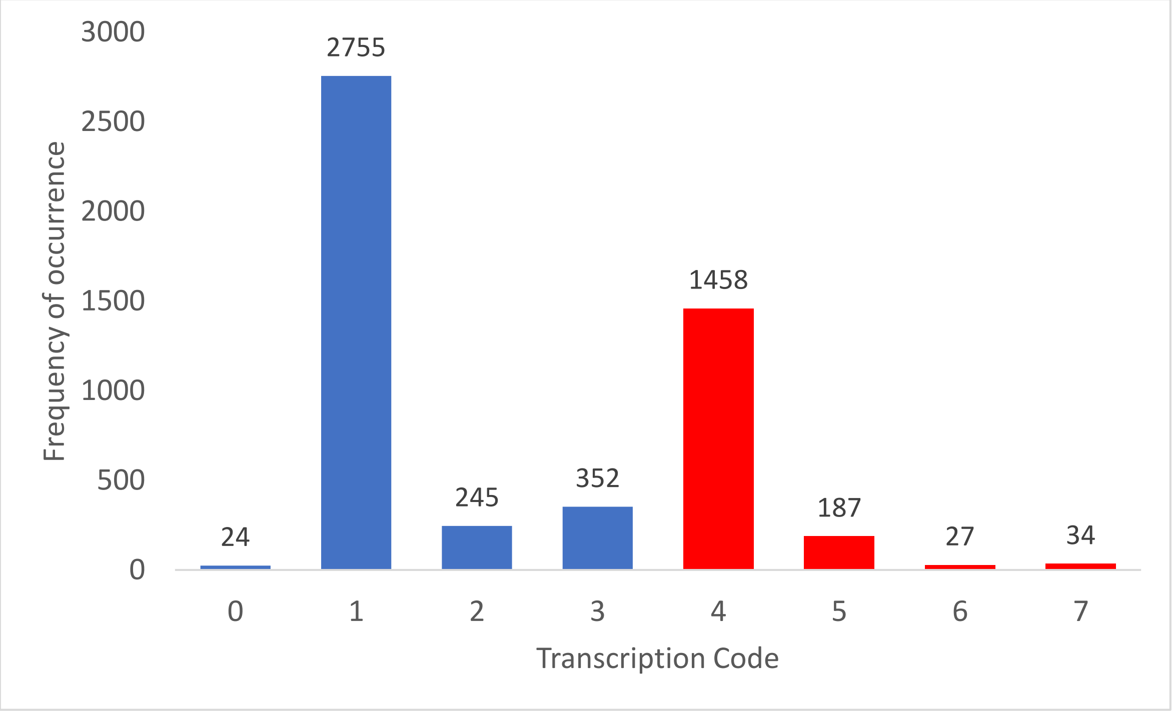
Association yielded far greater effects as differing levels were associated with starkly different odds ratios of correctly answering a connected speech item. With GLMM outputs, one level of any multilevel variable is used as a default comparison with the other levels; the default, selected by alphabetical order, is given the value 1 and not displayed in the output table. For the present output, association level "equal" was the default. "Same" had an odds ratio of 2.3, meaning that performing well in equal associations meant greater odds of performing well with same associations. The lower odds ratio for opposite associations indicated that the opposite task was not only difficult, but that performing well there was a strong performer of efficacious performance on other associations.

[55] Odds ratios of 1 suggest no effect; odds ratios < 1 indicate lower odds of an event; odds ratios > 1 indicate greater odds of an event.

Q1e. To what extent do known confounds (spelling and vocabulary knowledge) affect listener performance? Transcription and vocabulary knowledge were anticipated confounds which required monitoring. This section investigates and transparentizes these elements.

Transcription. The selection of transcription as a response type for Experiment 2 helped provide a more direct link to perception than was afforded in Experiment 1, but did so at the cost of potential confounds such as spelling and memory. It was thus important to monitor transcriptions for such confounds, and this was achieved through coding (see Experiment 2 Methods). Figure 11 summarises transcriptions of the Diverse Sentences task. Diverse Sentences was selected as an exemplar task as it was the principal target for comparison with bVt tasks and reflected the most numerous, diverse, and lengthy transcriptions among all sentence tasks.

Figure 11. Diverse Sentences transcription codes and their frequencies



Note. Codes 0-3 marked as correct; 4-7 incorrect.

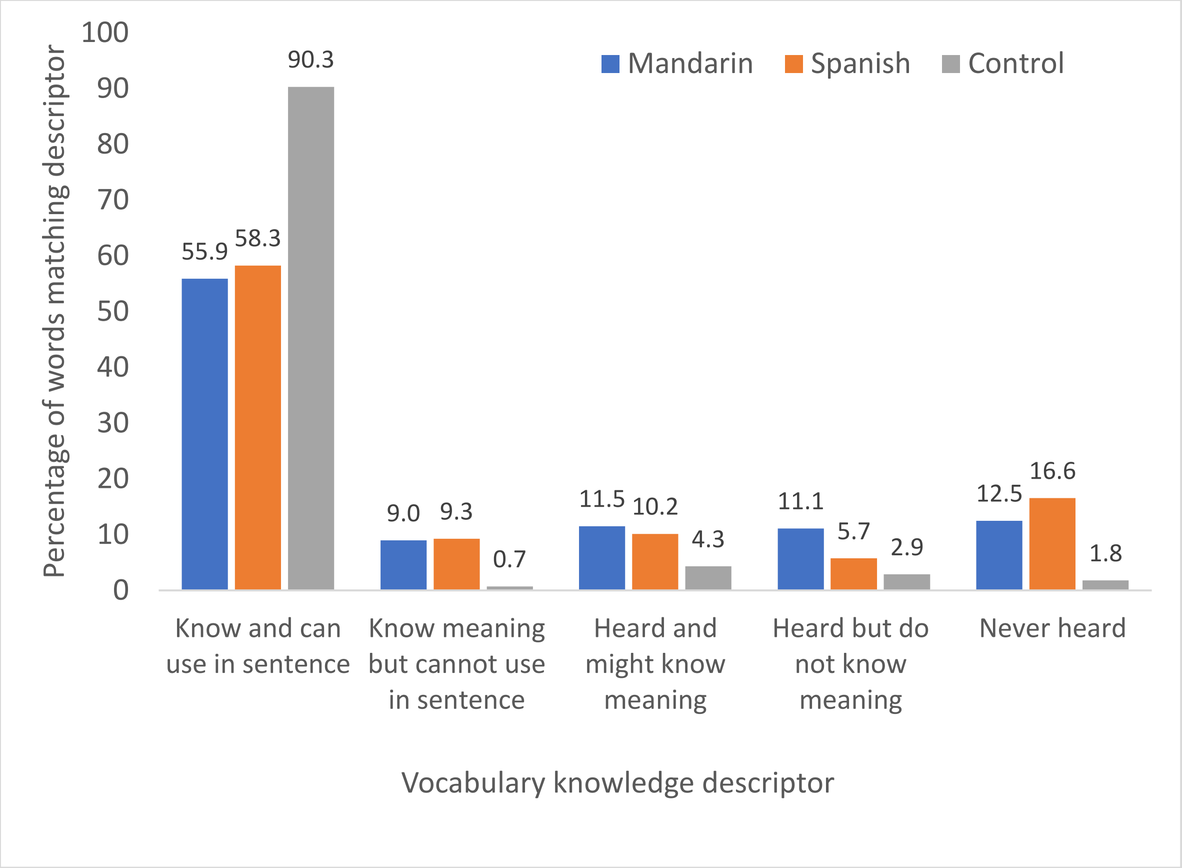
Legend: 0 = ambiguous transcription that has been given benefit of the doubt; 1 = correct vowel, word, and spelling; 2 = unambiguously correct vowel and word, incorrect spelling; 3 = correct vowel, incorrect word; 4 = opposite minimal pair perceived; 5 = vowel and word other than the minimal pair perceived; 6 = ambiguous or uninterpretable transcription; 7 = incomplete transcription, word with the target vowel not transcribed.

Transcription code frequencies (codes 1-5) suggest it was typically clear when the participant perceived the incorrect or incorrect word in the target vowel pair, supporting the efficacy of transcription for higher functioning L2 speakers. Participants were able to unambiguously convey the correct vowel even in instances where the word was misspelled (code 2) or, as displayed through orthography, a different word was heard (code 3). Codes 0, 6 and 7 reflect irregular findings which would be problematic if inordinately frequent. The primary concerns regarding transcription were spelling (subsumed by codes 0 and 6) and memory (principally subsumed in code 7). These codes cumulatively reflected less than 2% of the data. Data with these codes were retained rather than excluded to reflect concomitant artifacts of real-world perception.

Given the above, the inclusion of transcription as a response choice does not appear to have been inordinately problematic for memory or spelling.

Vocabulary knowledge. Knowledge of each minimally paired word was measured with a self-report survey after the listening tasks had been completed. Figure 12 summarises the results.

Figure 12. Summary of participant familiarity with target words



The bulk of target words were reported as "known" across all language groups; however, varying degrees of familiarity were shown at all points of the scale.

Linking word familiarity to performance (percent correct), of the total observed scores marked 1-4 for familiarity, 96.3% were correct. The "know and can use in a sentence" observations (n = 504) yielded 98.2% correct. There was a clear, descriptive discrepancy in performance shown between the extremes of familiarity. Focusing on the largest group, Mandarin, there were 310 observations for "never heard" with a mean of 42%. Of the 1386 observations for "heard and can use in a sentence", the mean was 72%. Though a discrepancy exists in performance with known versus never heard words, relatively few words (less than 17%) had not been encountered by L2 participants.

As stated in the generalised linear mixed model, there is likely an overlap with word familiarity and context association. Separating the two was beyond the scope of the study, but will be revisited in Limitations.

Following individual word familiarity summaries, modes for each minimal pair were calculated for the Mandarin and Spanish groups (cumulatively) for the 24 vowel pairs in Diverse Sentences. Results are summarised in Table 29.

Table 29. Mode vocabulary knowledge across /i, ɪ/ and /ɛ, æ/ vowel pairs in Diverse Sentences

/i, ɪ/	Word 1 Mode	Word 2 Mode	/ɛ, æ/	Word 1 Mode	Word 2 Mode
Pair 1	5	5	Pair 13	1	1
Pair 2	5	5	Pair 14	-	-
Pair 3	5	5	Pair 15	5	1
Pair 4	5	5	Pair 16	5	5
Pair 5	5	5	Pair 17	5	5
Pair 6	5	5	Pair 18	-	-
Pair 7	5	5	Pair 19	5	5
Pair 8	5	5	Pair 20	5	5
Pair 9	-	-	Pair 21	5	5
Pair 10	5	5	Pair 22	5	5
Pair 11	-	-	Pair 23	5	5
Pair 12	1	1	Pair 24	1	1

Note. hyphen (-) indicates target word was a proper noun (two per vowel pair); proper nouns were excluded from familiarity survey.

While a degree of discrepancies in familiarity is justifiable—it was impossible to perfectly control for idiosyncratic familiarities across individuals—modes suggested generally well-matched vowel pairs. All minimal pairs except Pair 15 had matching modes. Due to the overlap in familiarity and association, it was expected that for Pair 15, Word 1 would be the word most associated with the context, and Word 2 would be most difficult and discriminating. Results support expectations as 82% (*SD* = 39) of participants responded correctly for Word 1 while 59% (*SD* = 49) responded correctly for Word 2.

Vocabulary knowledge results indicate that the chosen words in the experiment were typically known by participants and word pairs were largely matched for frequency.

Q2. What are L2 participants' perceptions of using diverse stimuli versus traditional stimuli? How does L2 participant experience (self-reported task load, subjective experience identified through open-ended question) with diverse sentence prompts compare with their experience with bVt prompts?

Q2a. What was the perceived task load (mental, physical, temporal, predicted performance, effort) of sentential prompts compared with bVt prompts?

Table 30. Mandarin mean NASA Task Load Index (NASA TLX) subscale scores across tasks

	NASA TLX Subscale					
Task	MD	PD	TD	P	E	F
Oddity	11.1 (4.0)	7.3 (4.5)	9.0 (3.2)	10.0 (3.8)	9.6 (4.0)	8.2 (5.4)
bVt Identification	8.5 (4.4)	7.3 (4.5)	8.4 (3.4)	10.5 (4.0)	7.1 (3.9)	7.7 (3.9)
Travel Agent	10.3 (3.7)	9.7 (4.4)	8.9 (3.33)	9.0 (3.3)	10.5 (4.1)	9.9 (5.6)
Question & Answer	12.2 (3.8)	10.8 (4.3)	9.9 (3.2)	8.1 (3.2)	12.3 (3.9)	10.6 (5.7)
Diverse Sentences	13.6 (3.6)	13.6 (3.6)	10.7 (3.6)	8.1 (3.0)	13.3 (3.8)	12.4 (4.3)

MD = Mental Demand; PD = Physical Demand; TD = Temporal Demand; *P* = Performance; E = Effort; F = Frustration. Values had maximum range of 0-20.

Table 31. Spanish mean NASA Task Load Index (NASA TLX) subscale scores across tasks

	NASA TLX Subscale					
Task	MD	PD	TD	P	E	F
Oddity	11.6 (5.7)	5.0 (4.9)	8.4 (5.1)	12.0 (4.0)	10.8 (5.6)	9.5 (5.7)
bVt Identification	12.2 (4.7)	10.5 (5.5)	10.5 (4.9)	9.5 (4.2)	7.2 (5.9)	9.0 (3.7)
Travel Agent	10.1 (6.3)	5.6 (4.1)	7.1 (3.8)	10.6 (4.8)	10.6 (4.5)	10.2 (5.9)
Question & Answer	11.4 (4.9)	8.9 (5.5)	9.8 (3.7)	10.6 (4.5)	13.5 (3.2)	10.2 (5.2)
Diverse Sentences	15.3 (4.1)	12.0 (5.8)	13.2 (4.9)	10.1 (4.6)	14.8 (4.2)	10.9 (6.5)

MD = Mental Demand; PD = Physical Demand; TD = Temporal Demand; *P* = Performance; E = Effort; F = Frustration. Values had maximum range of 0-20.

Table 32. Control mean NASA Task Load Index (NASA TLX) subscale scores across tasks

	NASA TLX Subscale					
Task	MD	PD	TD	P	E	F

Oddity	10.6 (5.68)	8.1 (5.24)	8.3 (4.31)	14.6 (4.54)	11.7 (5.62)	8.0 (4.58)
bVt Identification	6.3 (5.82)	5.9 (5.84)	5.1 (5.27)	17.3 (3.82)	6.1 (5.37)	6.1 (5.05)
Travel Agent	11.1 (6.09)	4.1 (3.63)	6.4 (5.03)	13.3 (3.73)	11.3 (5.06)	7.9 (5.11)
Question & Answer	12.0 (3.96)	6.1 (6.96)	8.9 (4.26)	12.6 (2.51)	12.0 (4.20)	9.9 (6.18)
Diverse Sentences	11.7 (5.68)	6.4 (7.30)	7.9 (4.41)	14.4 (2.23)	11.9 (3.76)	8.6 (5.50)

MD = Mental Demand; PD = Physical Demand; TD = Temporal Demand; *P* = Performance; E = Effort; F = Frustration. Values had maximum range of 0-20.

Relationships between subscales and performance were considered using Spearman's rho. Subscales other than (perceived) Performance are expected to negatively correlate with score: the greater the mental, physical, or temporal demand or effort, the more challenging to obtain high performance levels. Participants' perceived (how they expected they performed) and observed (how they statistically performed) performance scores are expected to positively correlate.

bVt Oddity. Mandarin was the only group to display a statistically significant correlation between Oddity task performance and perceived workload. The Mandarin perceived and actual performance was positively correlated at .63 (p < .001). Findings were non-significant for the Spanish and Control groups (p > .05). Reported mental demand, physical demand, temporal demand, effort, and frustration were not associated with participant performance.

bVt Identification. For bVt Identification, Mandarin again was the only group to show a statistically significant finding for correlation between task load and performance. Mental demand and performance on the bVt identification task was negatively and significantly corelated (r_s = -.43, p = .02).

Travel Agent. Spanish perceived performance and actual performance were the only statistically significant finding for the Travel Agent task (r_s = .63, p < .05). No significant findings were discovered in the Mandarin and Control groups.

Question and Answer. For the Mandarin group, a statistically significant negative correlation between Mental Demand and performance (r_s = -.48, p < .01) and between Effort and performance (r_s = -.42, p < .05) was found. No other statistically significant correlations were identified.

Diverse Sentences. The Mandarin group performed as would be expected across the subscales. Mental Demand correlated positively with all other demand measures: Physical (r_s = .53, p < .01), Temporal (r_s = .63, p < .01), Effort (r_s = .6, p < .01), and Frustration (r_s = .43, p < .05). Performance was negatively correlated with demand measures for Mental (r_s = -.39, p < .05), Temporal, Effort, and Frustration, but not Physical.

For the Spanish group, the subscales Mental Demand, Effort, and Frustration were positively correlated Mental Demand correlated with Effort (r_s = .86, p < .01), while Effort was positively correlated with Frustration (r_s = .6, p < .05). Correlations were non-significant between subscales and observed performance.

The Mandarin group's physical (r_s = -.38, p < .05) and temporal (r_s = -.41, p < .05) demand correlated significantly with performance. No other correlations were found for the subscales among other groups.

Overall, findings which were significant aligned with expectations: subscales were negatively correlated while perceived performance was positively correlated. The Mandarin group's reported perceptions had modest success relating to performance, while the Spanish and Control groups performed well or poorly regardless of demand or perceived performance. Only once did the Spanish group predict performance at a significant level. Given only one or two subscales correlated with performance for each task, and never consistently, correlations should be met with a level of scepticism as there is a strong chance of type 1 error.

Q2b. How does L2 subjective experience (as identified through an open-ended question) with sentence prompts compare with their experience with bVt prompts?

I address this question by cumulatively displaying and discussing results from the open-ended question. Points of interest are further explicated in the formal Discussion section. To preserve participant privacy, I have pseudonymised responses where quotations are given.

Preliminaries (intercoder reliability). After Coder 2 had coded all participant responses, I compared his coding with my own in NVivo using the Code Comparison query and found inconsistencies that needed to be addressed. Overall results showed substantial agreement, yet individual code indices were not all satisfactory. Positive affect showed a negative kappa value (K = -.0158), suggesting disagreement between coders. Investigating further, Coder 2 had coded positive affect when participants referred positively to the experiment in general. These codes and data were removed as they did not conform with the indicated coding conventions. Additionally, I had missed coding one instance of positive affect. Coder 2 coded the following as positive affect:

> Even though the whole sentences provided some information, this was not enough for me. I loved the sentence transcription, but I have plenty of problems with my spelling, and that frustrated me a lot.

Though the coder included the first sentence for context, I felt the context was not suitable to explain or describe the positive affect (which was directed at liking the transcription despite being poor at spelling). Instead, I coded: “I loved the sentence transcription” as positive affect.

Re-running NVivo’s coding comparison for a revised reliability measure after coding corrections, the overall kappa coefficient and percentage of absolute agreement between coders (Table 33) was sufficient to conclude substantial overall inter-coder agreement (Landis & Koch, 1977).

Table 33. Kappa agreement coefficients, percentage of absolute agreement, and frequency of coded category in the dataset for Experiment 2 codes

Code	Kappa coefficient	Percentage of absolute agreement	Total occurrences
Attention\focus and general attention	0.5758	93.39	8
Attention\confusion	0.3930	87.30	12
Attention\fatigue	0.9982	99.99	4
Attention\memory	0.9941	99.97	2
Emotion\negative affect	0.5609	94.06	7
Emotion\positive affect	0.4378	97.91	2
General Cognition	0.5715	81.03	21
Perceived difficulty	0.6438	82.42	39
Strategies	0.4846	85.00	7
Prompt\bVt oddity	0.8244	95.37	19
Prompt\bVt transcription	0.5799	90.88	13
Prompt\question and answer	0.9649	99.19	14
Prompt\diverse sentences	0.8131	92.47	22
Prompt\travel agent	0.8962	96.36	22
Transcription	0.6625	90.61	18
Average	0.6934	92.40	-

In the next sections, I report and discuss the qualitative data generated from participant responses. I have broken this section into subsections to facilitate reading; however, findings could not always be fully compartmentalised under a single heading, and occasionally there was an overlap. For instance, in explaining the difficulty of isolated speech tasks (discussed first), a participant may compare it with a sentence task (discussed subsequently). In such a case, I included the comparison quote about the sentence task for reference and where possible, I used a separate, similar quote under the connected speech task subheading.

Task design, perceived difficulty, and performance. I designed the prompts to steadily increase in difficulty, though the results did not consistently match this intention. The predicted perceived difficulty, based on the initial design and arranged from least to most difficult, was bVt tasks (isolated, monosyllabic words), Travel Agent (participants had to listen for a single, sentence final word in a predictable task; participants type a single word), Question and Answer (participants' attention is focused to a specific piece of information; participants type a single word or phrase), and Diverse Sentences (no attention is drawn to target words; the entire sentence must be transcribed). Qualitative results showed mixed perceptions of difficulty. Some participants aligned with predictions,

> [bVt Oddity] Quite easy, the least mental challenging. [bVt Transcription] Relatively easy as well; Mistyped a little. [Travel Agent] It felt quite hard for me because I am not familiar with the names of these places. The difficulty for me [with Question and Answer] depended on the question - if it contains words I was not very familiar of then I would find it hard, otherwise it's ok. I find [Diverse Sentences] to be the hardest for me because there were many words I was unsure of how to spell. The speed sometimes got a bit fast for me as well. (Oliver)

As shown, Oliver's perceived difficulty matched the graded complexity of the prompt types: the easier prompts were those with least complexity while the most difficult were the prompts with most difficulty. The manual requirements of the task (i.e., typing and spelling) were also indicated as challenging, something expected from methodological switch to transcription from the 2-alternative forced choice task.

Other participants, discussed subsequently, identified different prompts as most difficult or easy. Unexpectedly, prompts where participants were able to focus strictly on isolated words (bVt) or single words in a sentence final position (Travel Agent) were identified as "most difficult" more regularly than Diverse Sentences.

In the Experiment 1 open-ended item Discussion, I posited that embedding target words in sentences for listening prompts may engage participants' attention or mitigate fatigue. In Experiment 2, there was some evidence supports this position, as Victoria stated, "got a little bit tired when [completing Diverse Sentences], but found it was easier to finish with context". For Victoria, context facilitated their completion despite the length of the task.

Particularly for participants who had difficulty distinguishing between target vowels, I expected and interplay between perception and lived experiences. Each participant came to the experiment with unpredictable, idiosyncratic associations which influenced how easy or difficult a sentence prompt would be for them. If a participant associated one word in a minimal pair with the sentence more than the other word from the minimal pair, the task would be easy. The word in the minimal pair that was not associated with the context would be more difficult. One of the participants, Tracey, explained that her lived experiences influenced how context impacted her perception in a connected speech task, "The 'slipping on the floor' one was tricky since in Chile, in public hospitals, some patients actually sleep on the floor when there are no available rooms so I related it to my previous experience". Regardless of whether the word was sleeping or slipping, Tracey heard the sentence as, "Patients keep sleeping on the floors" and had not considered the alternative word, "slipping" until seeing it in the familiarity and association surveys. This is consistent with her semantic network, where sleeping would be more readily accessible than slipping given the sentence context.

Task-related findings. Participants principally reported degrees of difficulty and related contributing factors. Similar to Experiment 1, there was no consensus regarding whether bVt or sentences generated the most challenging prompts.

Isolated speech prompts. Between bVt Oddity and Transcription, Oddity was often considered the easier of the two. Some participants labelled bVt Oddity as “fine”, “quite easy” and “the easiest” prompt type (overall). For those who explained their responses, reasons included being the least mentally challenging and “easy because comparison effect” (Christina). The “comparison effect” was in reference to having three isolated words in succession, allowing Christina to more readily identify qualitative differences between vowels.

Participant effects (i.e., where participants attempt to interpret the experiment and adjust their responses accordingly) were a factor for at least one participant in the oddity task. Extending the ability to immediately compare minimal pairs or hear multiple iterations of the same word by different speakers, Fraser found merit in what may be considered inadvertent high variability phonetic training, “this section wasn't too difficult to notice some differences between bet and bed, because of the different accents. I feel more familiar with the British accent now”. Though I valued the participant’s positive disposition toward the study, she had incorrectly identified the target I was testing. “Bet” and “bed” were neither minimally paired nor examined, and the participant was perceiving differences (false positives) which were absent in the design. Based on the participant’s transcriptions, she tended to perceive /b/ as /p/, and /t/ as /d/. This may have distracted her from focusing on the target vowels as, for the mid-low vowel pair, she received 81% for bVt Oddity (the task she acknowledged as helping identify distinctions), but 100% for bVt Transcription.

The bVt Transcription task elicited a range of participant responses, such as “easier than the discrimination task because I only needed to focus on a single word” (Dustin), “A little bit more difficult when I was writing” (Giselle), and, “I feel I got a good score for bVt Oddity. bVt Transcription was very difficult. I didn't know what they're saying” (Danna).

Several participants indicated that the isolated prompts were the most challenging. Dustin explained, “This was difficult! I haven't done anything like it before. I thought I could discriminate vowels and sounds but it seems I need to keep learning haha…The [bVt Oddity and bVt Transcription] tasks were the hardest”. The challenge of the tasks led to a degree of anxiety for some, as Danna indicated, “I feel anxious about bVt Oddity, almost can't figure out the differences”. Danna’s difficulty-induced anxiousness was evident in her performance on the high vowel pair Oddity task as she scored lower (obtaining 44%) with this task than on any other task. The difficulty, however, was a function of the vowel pair rather than the task itself as she performed at 84% in Oddity task with the mid-low vowel pair.

Various explanations were given for the perceived difficulty of bVt prompts, including maintaining focus, “easy at first but hard to distinguish when I heard a lot of them” (Oliver); phonologically-based insight, “Sometimes it is really difficult for me to discriminate some vowels following b” (Nigel); and most commonly, an absence of context, “more difficult because only able to rely on vowel itself and surrounding phonetic cues” (Brianna). Along with an absence of context, Lana stated, “it was kind of frustrating because most of the time I got words from the context of the sentence, so the b-vowel-t discrimination tasks was like... ok whatever!” Lana further vocalised “not bothering” with the bVt prompts and not seriously responding by listening to isolated speech prompts. Unfortunately, on the basis of Lana’s demonstrably biased temperament toward the bVt prompts, stated disregard for responding to the task, and correspondingly irregular results, I omitted her quantitative data from analysis. Her qualitative response remains a valuable consideration. It acts as a counterpoint to perspectives which view diversity (beyond fixed, isolated words) in speech perception assessment as distracting.

Connected speech prompts. Exploring the potential use of sentence prompts for listening perception assessment was the purpose of this research, and here I attend to participants' feedback on sentence-based prompts (Travel Agent, Question and Answer, and Diverse Sentences).

Continuing the discussion from bVt, participants viewed sentential context as helpful, but with various factors increasing difficulty.

> [I]t was really difficult, for me at least, to discriminate some words without the broader context in which they were used. Even though the whole sentences provided some information, this was not enough for me. I loved the sentence transcription, but I have plenty of problems with my spelling, and that frustrated me a lot. (Murray)

Murray appreciated the additional information of the sentence context compared with the isolated prompts, but spelling was a challenge. Other participants noted that, "as a non-native speaker, I was quite unfamiliar with spellings of locations" (Jane), that they "can not spell the words with confidence" (Ben), that there may be "different ways to spell the names of a place" (Brianna), or that they were "struggling about the spelling and feel frustrate about the names of the places" (Liam).

Another challenge, familiarity with words, was often stated. As Brianna mentioned, "[Sentence prompts] are quite difficult because some words are new to me and sometimes I cannot recognize and understand what they say". Lack of familiarity was characteristic of the Travel Agent task. Recalling that the Travel Agent task mirrored Experiment 1's Directions Task in task design (listening for location names which participants were often unfamiliar with), it is fitting that the feedback also corresponded. Several participants linked location names with expressions equating to "did not know" and "difficult". As Christina explained, "Personally, I found the Travel agent apart is the most difficult because I do not know the place names". Fraser elaborated that the lack of familiarity with place names was detrimental to his effort and performance, "The unfamiliar place name to some extent impedes my answering. Although I know the distinction of vowel sounds is what I should pay attention to, I did not put much effort to distinguishing them".

Though most references to familiarity targeted the Travel Agent task, one participant indicated a connection between familiarity and the mental demand of the other sentence-based tasks, “I find it more challenging in [Question and Answer] and [Diverse Sentences] sections and they are more mentally demanding, especially when I heard the word which i don't understand” (Alissa). Though Alissa did not elaborate on mental demand, she may have meant memory or attention, as indicated by Fraser, “In the Diverse Sentences section, I was really distracted by remembering the whole sentence and it needs much efforts than other sections”. I will discuss the interaction between effort and performance later in this section.

Resolving uncertainty. Unfamiliarity (particularly with location names) and uncertainty for some may have led to guessing-based strategies in responding to prompts. Aaron described transcribing the place he considered potentially real rather than the place he heard, “[A] few of these places I could identify or sounded real and others did not. More than listening to the sounds, I was listening for real-sounding places”. Similarly, Stacy explained,

> In the booking task, I thought the places were in the US or Canada, because that’s where the experimenter is from. So I changed my answer sometimes. Then I thought, we’re studying in the UK, so maybe the places are from the UK, so I tried to answer thinking of UK places.

Participants described responding to prompts that they were unsure in accordance with instructions (e.g., spell the word how it sounds), uncertainty, or discomfort. Dustin expressed, “Regarding answering questions and diverse sentences, i have mostly written the words as they sound to me, even so it sometimes didn't make much sense in the sentence”. Katia indicated feeling “a bit unsure about my answers” because “sometimes I thought the sentence didn't make sense even I wrote down every single word (it seemed that the sentence did not follow a correct logic)”, and Aaron “felt uncomfortable just guessing”.

Reference to familiarity extended to speaker characteristics. For Gerrard, “The most challenging one is the Travel Agent because the names of the place can vary a lot, depending on the accent or dialect of the speaker”, and for Matilda, “sometimes the accent made the words' phonetics difficult to grasp”. To interpret these excerpts, I must first address participants’ use of the words “accent” and “dialect”. As vowel changes and productions were controlled for[56], and talkers spoke a similar English language variety, I would not expect one talker’s accent to help identify individual sounds or words more than another talker’s accent. If accent did play a role, the effect would be uniform throughout the experiment and the same for all talkers. Further, it would not handicap or benefit a particular task. However, participants may have used the terms “dialect” and “accent” to mean “talker characteristics”, and as discussed in Methods, familiarity with a talker’s voice has been empirically demonstrated to impact listening performance in speech perception studies. Voice specificity effects enable participants to perform better on speech perception items which include previously heard voices (Papesh et al., 2016), and voice specific priming—where listeners adjust perception based on talker idiosyncrasies—helps listeners correct talker errors (Goldinger et al., 1989). Again, these effects would hold for both sentences and isolated words, and the same talkers were used throughout the experiment. Further, participants were given practice items which included the same talkers as the experimental items to help facilitate perception. Although I am satisfied that accent, as I have defined it, did not unduly impact participant performance, one Talker was statistically more difficult than his counterparts and individual participants performed better with specific talkers.

I have discussed the effect of not knowing a word; however, familiarity did not always help participants when they knew words, as illustrated by Giselle, “I felt a bit frustrated especially when completing tasks 4 [Question and Answer] and 5 [Diverse Sentences] because I couldn't recall words that I know…the sound of them didn't trigger my memory”. Giselle did not consider certain (known) words until seeing them after the listening tasks in the vocabulary knowledge and association sections. Giselle’s L1 and sentential context can help explain her response. She has difficulty perceiving the difference between target vowels, so she relies on context to disambiguate.

[56] Minimal pairs which contained target vowels in American English but not British English (e.g., “command” and “commend”) were omitted and talkers who merged /æ/ and /ɛ/ were excluded.

Presumably, she associated one word in the minimal pair more strongly with the context, and this precluded her from considering the cognate word in the minimal pair. However, context did not invariably dominate perception, as seen with Victoria, “Although I knew some words very well…(like mills and fill), I did not think to write them even though they make more sense in the context than the ones I actually wrote”.

Discussion

This study supports and extends the findings of Experiment 1. Both studies displayed strong internal consistency and showed limited generalisability of bVt prompts to sentential prompt types. Experiment 2 quantified the added task load of increased prompt variability (and increased item numbers), but there was little correlation with actual performance. A series of correlations yielded few significant results, suggesting that for the experiment, increases in mental, physical (i.e., typing), and temporal demands were not enough to meaningfully affect performance, nor was increased frustration.

A primary aim of Experiment 2 was to identify the effects of association on accurately perceiving target vowels in sentences. Connecting association results with performance uncovered a key discriminating factor between high and low performers: the ability to accurately perceive a word of a minimal pair despite it not being the word of the minimal pair most strongly associated with the sentential context. The strongest performers perceived the vowels accurately regardless of the context.

Many participants performed well in Diverse Sentences when association was same or equal, yet there was a precipitous drop in performance when the context was opposite to what participants associate with the target word. This steep drop is not exhibited by high performers in the Mandarin or Spanish groups, nor in the Control, hence something is occurring where the majority of participants are being affected by something that others are not. Such results from the Mandarin group may be explained by an imbalance of top-down and bottom-up processing. The Mandarin group exhibited a basic, bottom-up capability to distinguish between the high vowels (/i, ɪ/). This was demonstrated by strong performance where association was equal between minimal pairs for a sentential context. Alternatively stated, when neither word was associated more strongly with the sentence context, the Mandarin participants tended to accurately identify which word was said by the talker. However, Mandarin group performance went from 79% to 39% when the sentence contained the word in the minimal pair that was not associated as strongly with the context. Because it was not associated with the context, participants perceived, or at least transcribed, the word that was. As the context rather than the segment was used to interpret the sentence in such cases, it appears top-down processing of the sentence context is predominant over bottom-up processing of the vowel, modulating participants' perception.

As evidenced by the higher performers and within the parameters of this study, there is a level of vowel perception which is largely robust to the top-down dominance we see affecting lower performers. It appears higher performers are better equipped to balance top-down and bottom-up processes, effectively utilising each to interpret a given utterance. Consequently, it is possible that listeners who are performing well with neutral prompts but poorly with "opposite" prompts are exhibiting an intermediate step in perceptual acquisition, one where the bottom-up process of vowel discrimination (for target vowels) is present, but fledgling[57]. The differing levels of perception suggest an incomplete functional perception, one dominated by top-down processing.

Strong performance with opposite associations in Diverse Sentences suggested strong performance across all tasks[58]. This suggests an implicational hierarchy: if listeners can consistently perceive a minimally paired word that they do not generally associate with a given context, they can perceive the distinction in words in equal and same contexts. Further, listeners can readily perceive

[57] Presumably, there would be less need to train these participants (those who perform well with equal prompts but poorly with opposite) with traditional prompts such as bVt; they can perceive the difference between sounds effectively in limited contexts, but do not possess a functional level of perception.

[58] Though correlations with oddity performance across tasks were moderate to strong, they were tempered as obtaining a low score in opposite associations did not translate to low scores in other measures (e.g., isolated speech tasks were performed well on even by those who did poorly in oddity).

target words in isolation. For perceiving the distinction between target vowels, the hierarchical chain may be broadly schematised,

> connected speech with opposite association > connected speech with equal association > isolated speech (identification and discrimination) and connected speech with same association

In the above schematic, there is a hierarchical performance chain, with ">" indicating that strong performance on the task to the left of the symbol suggests strong performance on the task to the right of the symbol.

In comparison to the Diverse Sentence prompts, bVt prompts, though reliable, may limit interpretations of performance and claims that may be made from them. The bVt prompts are not well-suited to reflect different levels of difficulty or learner attainment. Further, performance on the bVt prompts was not shown to be a strong predictor of performance with connected speech prompts. The bVt prompts do, however, offer an initial step in uncovering whether a listener is able to differentiate between vowels. If the low rung of an implicational hierarchy, performing poorly with bVt tasks would likely preclude a listener from performing well on more phonologically and sententially diverse tasks, particularly when there is little context (e.g., where association was "same" in the Diverse Sentence task). Such would be a meaningful diagnostic, particularly at lower levels.

The open response type, as the closed response type in Experiment 1, had high internal consistency. Less than 2% of the transcribed data was identified as uninterpretable for Diverse Sentences, whether from ambiguous spelling, partial responses, or absent responses. Such results support the use of coded transcription for future sentence perception tasks. Employing the transcription task did have a qualitative cost, however, as many participants displayed uncertainty about their spelling. It is possible that negative affect may compound the difficulty of the tasks, particularly for lower-level learners (Asseburg & Frey, 2013).

Various challenges were mentioned in the open-ended response, and the challenges from spelling, familiarity, memory, time, and pronunciation led to frustration for some participants. These elements may have influenced results, particularly for participants who had a hard time hearing the distinction between target vowels. Still, at the group level, participants who were predicted to do well based on PAM-L2 did perform well. First language speakers of English all performed well despite experiencing the same frustrating items, and the Spanish L1 performed significantly better with the front mid-low vowel pairs than the front high. This suggests that phonological perception may be, at least to the extent of the experiment, robust to fatigue or negative emotion in higher level listeners. There would presumably be a limit where these factors would negatively impact performance and further study would be warranted to explore those limits.

Results have theoretical and practical implications. For theory, results suggest there may be additional levels of acquisition we have yet to account for in the literature or in existing models of perceptual acquisition. PAM-L2 for instance, does not account for gradations in acquisition within each assimilation type, such as asymmetries between bottom-up and top-down processing. For instrument development, we may effectively employ familiarity and association to not only toggle difficulty, but identify areas which have hitherto been subsumed as a single, indivisible union: top-down processing as it directly relates to its bottom-up complement. A deeper understanding of participants by including listeners' target word associations with each sentence prompt, is provided. This research shows potential for added functionality in contemporary vowel perception assessments, accounting for both top-down and bottom-up processing and offering a means of assessing it through instrumentation.

General discussion (research synthesis)

Two studies investigated the effects of employing increasingly diverse listening materials for assessing vowel perception in adult L2 learners. The studies examined the suitability of less constrained listening items for diversifying the type of stimuli that could be used to reliably assess listeners' ability to differentiate between L2 vowels. Experiment 1 employed a two alternative, forced-choice design for identification tasks. Experiment 2 extended Experiment 1 with word association and familiarity tasks to explore the effect of association on assessing vowel perception in sentential contexts, and a task load survey was administered to identify the perceived workload of individual prompt types. Both studies complemented identification tasks with oddity tasks to assess discrimination. BVt and diverse prompt types were compared quantitatively and qualitatively. Following design practices of Bilger (1984), the studies assessed speech perception prompt (item) functioning through internal consistency (Cronbach's alpha), proportion correct, and the prompts' ability to discriminate between high and low performing participants. Predictions for participants who were expected to perform highly or poorly for target vowel discrimination were informed by the PAM-L2 framework, and prompt types were assessed based on degree of correspondence with PAM-L2 predictions. The research employed generalised linear mixed models to explore prompt-level complexity and investigate the extent to which performance with bVt prompts predicted performance with Diverse Sentence prompts. A thematic analysis was performed for each study on an open-ended item asking about participant experience.

In this section, results are summarised in relation to each other and discussed in reference to the broader literature. Primary emphasis is placed on the generalisability of bVt prompts, the suitability of more diverse prompt types for assessing L2 vowel perception, and the effect of association on vowel perception in sentences.

Generalisability of bVt prompts and implications for assessment design

Identifying the extent to which bVt generalises to connected speech is useful for researchers and assessment designers interested in defining the applicability of fixed frame prompts. Experiment 1 and 2's generalised linear mixed models revealed that bVt did not well predict performance with Diverse Sentences. Odds ratios were consistent for both studies and both vowel pairs, with a negligibly small, positive effect for predicting Diverse Sentence performance with bVt identification and discrimination tasks. Such results are congruent with claims that perception is context sensitive (e.g., Flege, 1995a; Thomson, 2012); however, these claims must be attenuated for diverse sentences as they are generated as a result of research designs which employ fixed consonantal frames.

For assessment design purposes where isolated prompts are assumed or required to generalise to connected speech, extending beyond a single fixed consonantal frame may be most appropriate as varied words (i.e., Experiment 1’s Diverse Words) better predicted connected speech performance than bVt. Relatedly, including both identification and discrimination tasks for analysis did not meaningfully add to the information gathered by one task alone. Further, the two task types predicted connected speech performance similarly, constituting redundancy. Consequently, in the absence of specific theoretical needs for including both tasks or for having one task over the other, either task (individually) would be equally suitable for making inferences about participant performance. More work is needed to uncover the generalisability of this claim, however, as the relationship between discrimination and identification tasks has been shown to vary based on segmental type, stimulus naturalness, stimulus context, and listener variables (Gerrits, 2001).

With the limited generalisability of bVt prompts established, their adherence to predictions will be explored in the next section concurrently with the suitability of diverse prompts.

Suitability of diverse prompt types

Foundational criteria of suitability were determined in alignment with Bilger (1984), focusing on internal consistency, proportion correct, and discriminability. This foundation was situated within an L2 context through PAM-L2 and extended to include interactivity between the participant and the assessment (Bachman & Palmer, 1996, 2010).

Internal consistency

Over the two experiments, bVt Oddity was the most internally consistent prompt type as indicated by Cronbach’s alpha. Independently, this is unsurprising given that internally consistency has been associated with homogeneity (Cronbach, 1951; Tang et al., 2014) or item interrelatedness

(Cho & Kim, 2014). From the first reported manuals discussing design methods for listening prompts (e.g., Fletcher, 1922; Fletcher 1929), fixed frame prompts have been employed specifically because the restricted consonantal and syllabic context constrains variability. Variance beyond the phone, to the extent possible, is eliminated purposely. In this sense, the consistency of bVt prompts was a confirmatory finding supporting the functioning of bVt as a prompt type. Confirmation aside, the more valuable insight gleaned from the studies is that diverse prompts, though to a lesser degree than bVt, remained sufficiently internally consistent within the confines of the two experiments. Specifically, items with added complexity and context performed sufficiently well to be incorporated in experimental research settings (Richardson et al., 1991), in learning materials (Fulcher, 2012), and summative assessments (Magimairaj et al., 2022).

The relative strength and weakness in internal consistency across prompt types was a finding of interest. Diverse word and sentential prompts were compared with bVt and a minimum criterion of α > 7. The bVt prompt type held the strongest internal consistency, but the non-adjusted (i.e., all items were used in the calculation) values for Diverse Words and Diverse Sentences were comfortably above the criterion. A relative lack of consistency was found where prompts had target words which were potentially less familiar to participants and were of a syllabicity associated with increased difficulty (i.e., multisyllabic words opposed to monosyllabic). This was the case for location tasks (Directions and Travel Agent) in Experiment 1 and 2.

The transcription response type displayed robust internal consistency. Experiment 2's Diverse Sentence prompts had half the items as Experiment 1, yet similar, if not stronger, internal consistency. After reducing the total item number of Experiment 2 Diverse Sentences by half again (totalling 24 items) for cross-prompt comparisons, values remained above the α > 7 criterion. This suggests that future studies may use far fewer Diverse Sentence prompts while remaining sufficiently internally consistent. Though vowel perception studies do not generally report Cronbach's alpha, Diverse Sentence prompts compared favourably with ranges reported in other speech perception research (e.g., Appezzato et al., 2018; Chen et al., 2018; Perdy et al., 2017;

Snowling et al., 2019). In sum, bVt was the most internally consistent prompt type, while Diverse Sentence prompts were sufficiently consistent despite having comparatively more complexity than bVt. Next, predictions are discussed in relation to prompt type.

Adherence to PAM-L2 predictions

Though identifying and reporting language group functioning was not the ultimate aim of this research, doing so was necessary to help establish proper prompt functioning and enable comparisons based on established theory. A major finding of the study was that listener performance with the Diverse Sentence prompts was strongly congruent with PAM-L2 predictions while performance with the bVt prompts was to a lesser degree. Larger sample sizes would bolster this claim; however the sample sizes of the present research are as large or larger than those in the most commonly cited PAM literature (e.g., Best et al., 2001; Best & McRoberts, 2003; So & Best, 2010; Tyler et al., 2014) [59].

Recalling the PAM-L2 categories that were assigned to each group, the Mandarin group was expected to assimilate /ɪ/ into /i/ due to vowel space proximity (Jia et al., 2006), but as a less good exemplar, making it "category goodness" (CG). The /ɛ, æ/ vowels were expected to assimilate into the same category (SC). Performance was therefore expected to be poorer for /ɛ, æ/ than /i, ɪ/. Korean had L1 analogues for /i/ and /ɛ/, but not /ɪ/ and /æ/. Proximity in vowel space suggests assimilation to the nearby categories (/i/ for /ɪ/ and /ɛ/ for /æ/), where the new vowels are less good exemplars of vowel categories (CG). The Spanish group was determined to be SC for /i, ɪ/ and TC for /ɛ, æ/; hence, participants were expected to perform decidedly better with /ɛ, æ/ than /i, ɪ/. For the control, performance was expected to be at or near ceiling for both vowel pairs. Whereas the control faired as expected across all tasks in Experiment 1 and 2 (at or near ceiling, no differences between vowel pairs), differences in L2 performance by task type yielded notable findings.

[59] These were the four most highly cited papers for PAM or PAM-L2 in Web of Science as of March 2022.

PAM-L2 predictions were mixed for bVt in both experiments. The Mandarin group was expected to perform better with /i, ɪ/ than /ɛ, æ/. In Experiment 1, the difference was not found to be significant for either bVt Oddity or bVt Identification. For Korean, it was expected that participants would, as a group, discriminate relatively poorly between /ɛ/ and /æ/, but better for /i, ɪ/. Poor in PAM-L2 has been described as up to 70% (Tyler et al., 2014, p. 4). This matched performance as the group cumulatively performed below the 70% range for /ɛ, æ/ for bVt. As with the Mandarin group, no statistically significant difference was found between Korean performance with /i, ɪ/ compared with /ɛ, æ/. In Experiment 2, the Mandarin group performed equally well on both vowel pairs in the bVt transcription task, but for oddity, performed opposite to expectations, where /ɛ, æ/ outperformed /i, ɪ/. Results aligned with predictions for the Spanish group, however, as the group was shown to perform better with /ɛ, æ/ than with /i, ɪ/ across all tasks.

The diverse prompt types, particularly Diverse Sentences, matched PAM-L2 predictions more consistently than did bVt. For the Experiment 1 Diverse Sentence prompt block, the Mandarin group obtained a greater accuracy score for the /i, ɪ/ pair than /ɛ, æ/. This was a statistically significant difference. Korean group performance mirrored the Mandarin group, performing significantly better with /i, ɪ/ than with /ɛ, æ/. The same was found for the Diverse Sentence block of prompts, where Mandarin again performed better with /i, ɪ/ than /ɛ, æ/.

Similar to bVt, the other diverse prompt types (Diverse Words, Directions, Travel Agent, Question and Answer) did not dependably reflect PAM-L2 predictions; differences were either significant in the direction of predictions (e.g. the Mandarin group with Directions in Experiment 1 and Travel Agent in Experiment 2) or non-significant (Koreans with Directions and Travel Agent). None, however, resulted in significant findings which were contra predictions.

Given that Diverse Sentences offered strict adherence to PAM-L2 predictions while other prompt types did not, the question became, “why?” The answer is likely a culmination of the prompt, the task, and the sample. For bVt, its strength—controlling for variability—may have also been its demerit, helping lead to irregularity in achieving predictions. Simply, it may have been too easy for the sample, both comparatively (with other prompt types) and statistically. When the proportion of participants who answer correctly on an item is high, the item's ability to discriminate (see Appendix for discrimination indices) between high and low performing participants is diminished (Sim & Rasiah, 2006). Vowel perception assessments are employed to identify whether a listener can distinguish between target vowels, but due to the ease of the bVt prompt, participants who are theoretically less able to differentiate between the vowels may still perform well. When added to the design choice of having two options to choose from for identification (Experiment 1) and four choices for oddity, the limited difficulty is exacerbated by increased chances of correct guessing. Resultantly, the information obtained by the bVt prompt about the participants' estimated abilities to differentiate between target vowels was mitigated, and results did not consistently reflect predictions nor strongly correlate with Diverse Sentences.

The isolated context also naturally helps the listener to directly attune to contrasting features. Thomson (2012, p. 234), for instance, employed isolated stimuli to help listeners attune to vowel-specific differences, reasoning, "learners require greater exposure to [difficult target segments] than they do for categories where a larger proportion of tokens are clearly dissimilar from any L1 category". The other information, he explains, is non useful and may prevent listeners from noticing distinctions. Thomson employs the same principle of restricted consonantal environments for both training and assessment (generalisation of training).

With the isolated prompt type designed to facilitate attention to specific, identifiable vowel characteristics in specific contexts, it depends on the listener to attune to distinguishing properties of the vowel. Intermediate and advanced learners are particularly equipped to perceive nuances in phonetic variation, both spectral (Lengeris, 2009) and suprasegmental (Wang & Munro, 1999). The present sample consists of L2 speakers of English who, as a group, are not only proficient at the C1 level, but navigate in linguistically advanced environments (i.e a university setting). Thus, it can be expected that they have effectively learned or developed strategies for circumnavigating ambiguous phones (Hasan, 2000), converging past experience, context, and linguistic knowledge (Donnelley, 1988[60]).

It is still unclear why or how the mid-low vowel pair outperformed expectations according to PAM-L2 predictions while the high vowel pair did not. With the high vowel pair, the temporal distinction was prominent in the present data set with bVt /i/ approximately 40% longer than bVt /ɪ/. Given the durational differences between front vowels, the predicted CG assimilation pattern for Mandarin and Korean groups, and the proficiency of the individual listeners, it appears irregular that /i, ɪ/ was not statistically significantly different than /ɛ, æ/. An initial explanation is that the study's samples of L2 listeners have learned to use spectral differences in the vowels or created a nascent vowel category for /æ/, enabling stronger performance with the /ɛ, æ/ contrast. However, PAM-L2 posits that new categories are most likely to form where one vowel in the contrast is a closer reflection of the L1 category than the other, as with CG, or when one or both are uncategorised. This entails that /ɛ, æ/ is least likely to form a category. Rather, it is more expected that /ɪ/ would form a new category because it is CG. To explain why the two vowel pairs would perform equally well within the bVt context (but not in Diverse Sentences), it appears the most likely answer is that it was cause by a type of ceiling effect.

The most glaring prediction violation occurred with Experiment 1's bVt Oddity task, where the Mandarin group performed better with the /ɛ, æ/ contrasts (SC) than with /i, ɪ/ (CG), contra to expectations. This finding is further curious because it does not appear to be due to guessing. The high scores compared to chance (chance scores for oddity were half that of identification, 25% to 50% respectively) combined with high internal consistency (α = .93 for /ɛ, æ/; α = .91 for /i, ɪ/) suggests guessing is not likely the culprit. The second study did not encounter the same result, and it may be that the additional stimulus (i.e., strings of four rather than three words) permitted added

[60] Donnelley was speaking about children with hearing problems rather than L2 learners, but as Hasan (2000, p. 139) explains, "Research on second-language listening comprehension draws on studies done on first-language learning (Anderson & Lynch, 1988; Devine, 1978, 1967; Duker, 1964; Dunkel, 1991;Keller, 1960). It can be said that much of the information we have about L2 listening comprehension is rooted in the work of first-language researchers".

opportunity for listeners to compare and contrast each token. As this is an isolated instance, it may most likely be an artifact of the study.

This section has indicated that the ease of prompts had a negative effect on satisfying predictions. It would be expected that if easiness affected predictions for the present study, similar findings would be presented elsewhere. Panning out to the literature at large, facility (easiness) may help explain why some studies have not met predictions while using similar prompts (e.g., de Jong, 1995, Reid et al, 2014) or have led to performance better than expected given the TC/UC > CG > SC assimilation hierarchy. Tyler et al. (2014), for example, investigated PAM-L2's application to vowels, reasoning that the bulk of research has focused on consonants. To illustrate their model, the researchers cited work from Polka (1995)[61], categorising L2 vowel contrasts through the lens of PAM-L2. Polka reported English L1 perception of the German vowel contrasts /u, y/ and /ʊ, ʏ/. Based on the study's categorisation (identification) results, Tyler et al. interpreted the /u, y/ contrast as CG and /ʊ, ʏ/ as UC. The discrimination scores, however, were at ceiling levels (98-100%) for the vowels labelled as CG, and "very good" (87%) for the vowels labelled as UC (p. 4). UC is akin to TC, and TC is akin to native perception. As UC was markedly lower than CG, and CG performed at ceiling levels, Polka's findings are contra to expectations. The researchers suggested the discrepancy may be due to the difference between consonants and vowels:

> [G]iven the shallower category boundaries and high within-category discrimination for vowels than consonants in categorical perception (Fry et al., 1962), SC assimilations might not occur for vowel contrasts, and absolute levels of discrimination could be higher for vowels than for consonants…the differences in discrimination performance that PAM predicts among SC, CG, TC, UC and UU assimilation pairs could be masked or overridden by the less categorical perception of vowel than consonant contrasts.

[61] The researchers cited two studies for the same purpose, the second study, Polka and Bohn (1996), utilised infants. As the current research addresses claims about adult perception, it was not appropriate to include this reference in the discussion.

Such intricate theorising could be alternatively yet equally summarised by something more parsimonious: the fixed consonantal frame prompt type was too reductive to reliably identify assimilation patterns. Physically distinct, but non-categorical differences (e.g., duration, F0) may have been responsible for the inordinately high results. By including additional variability in talker and phonological context, L2 listeners would be less able to focus on secondary cues to differentiate between vowels. Polka used naïve participants, but with more advanced learners, the additional contexts may have also enabled further refinement by identifying which prompts were most challenging or by finding that the limited frame generalises beyond an isolated, phonologically restricted context.

Though the present research offers support for the use of more diverse environments in vowel perception assessment designs, a disclaimer should be made that findings do not negate all uses of restricted contexts. Fixed consonantal frames can be a purposeful design choice for theoretical inquiry, such as for establishing perceptual thresholds (Nunn et al., 2019), investigating relative effects of individual formant frequencies (Hillenbrand et al., 1995), or in explicitly exploring individual phonetic contexts' effects on perception, such as SLM (Flege, 1995a) and SLM-r (Flege and Bohn, 2021). However, for establishing perceptual constancy or vowel assimilation—that the listeners' have assigned the vowel phonemically rather than phonetically (i.e., not simply a phone, but its variants or allophones)—prompts which can assess participants at a more nuanced level may be most suitable (see Association for further discussion).

Participant experience with fixed and diverse prompt types

Including participant experience and interactivity is a component of contemporary assessment literacy (Bachman & Palmer, 2010) which helped build on the speech perception design framework established by Bilger (1984). Participant experience was investigated with an experiment-final question in the two experiments, in addition to a task load survey (NASA TLX) at the end of each block of listening prompts for Experiment 2. Prompt difficulty was a key feature identified by participants in the NASA TLX and the experiment-concluding open-ended question. As expected, there was a general increase in task load corresponding to the increased complexity of prompt types. However, the heightened task load of the diverse prompts did not yield a systematic decrease in overall task performance as demonstrated by listener scores. In other words, listener performance was approximately equally good or poor regardless of demand. This indicates that within the parameters of the present study, listener perception was resistant to increased task demands, or that the increased demands were not sufficiently high to impact performance. Findings align with research showing that listening effort described by attentional or cognitive requirements may not consistently correlate with scoring (Strand et al., 2018). Presumably, mental and physical demands were not detrimental for the present studies' high intermediate-advanced English, university-level participants. How well vowel perception would remain robust for individuals of lower language proficiency or to demands beyond those applied in the present research would be a meaningful contribution to the literature.

In the open-ended question, both experiments showed mixed responses for bVt and the more diverse prompt types, particularly for sentences. Advantages reported for sentences included the benefit of being able to choose words from context (or being able to use context to facilitate guessing, as may be done in a real-world setting), increased time to process words, and increased focus. The isolated words were reported by some listeners to increase frustration and fatigue due to their repetitive nature. Contrastingly, some participants felt that context was only helpful if participants were familiar with the words or sentences, or that sometimes the sentence context would be misleading. Because both words in each minimal pair led to semantically and syntactically possible sentences, the opposing claims that context helps or misleads participants can be viewed as commensurately incorrect, or alternatively, equally correct. It would only help where a participant associated the key word (opposed to the opposite word in the minimal pair) with the sentence; conversely, it would only 'mislead' in situations where a listener did not associate the key word with the sentence with the sentence context (see Associations for explication of both situations).

As elicited by individual participants from the two studies, the question remains: does the added variability constitute a construct irrelevant distraction or is the added variability in sentential prompts relevant to the construct of vowel perception? Field (2004, p. 370) found, L2 participants "place more confidence in their pre-formed schema than in incoming data from the speech-stream", indicating that L2 participants would tend to follow context rather than the incoming speech signal. Hence, if the purpose of the vowel perception assessment is to examine vowel perception in a specific context for theoretical purposes, the use of sentence prompts could reasonably be viewed as leading to construct irrelevant variance; variance in scoring would occur from an interaction of bottom-up and top-down processes, not simply bottom-up (Field, 1999, 2004). To focus strictly on bottom-up in a single environment, a fixed consonantal frame such as bVt would be warranted. If attempting to offer a more functional view of vowel perception, one which investigates phonological rather than strictly phonetic perception and engages both bottom-up and top-down processes to approximate more natural speech processing conditions, the employment of diverse sentences could be justified.

Neither bVt nor more diverse sentence prompts were overwhelmingly preferred or indicated as preferential by participant responses. Therefore, regarding participant experience, sentence prompts appear to be at least as suitable as bVt prompts for assessing vowel perception. The effect of association on participants' vowel perception will now be addressed.

The effect of association on vowel perception in sentences

A significant contribution of the present study was introducing connected speech for investigating adult L2 vowel perception and exploring the effect of association. The interplay between association and vowel perception in sentences was explored in Experiment 2. In the experiment, when a listener heard the recording of a talker's utterance, associations were activated (Collins & Loftus, 1975), helping decipher what was heard and filling in any absence of information encountered with bottom-up processing (Field, 1999), such as disambiguating what may be perceived by some L2 listeners as homophonous words. This was explored in three conditions: when

Word A was more associated with the sentence than Word B and Word A was the key (association was “same”); when Word A was more associated with the sentence than Word B and Word B was the key (association was “opposite”); and when neither word was associated more strongly with the sentence (association was “equal”).

Same associations

With *same* associations, high scores were found where low scores were predicted (i.e., Spanish /i, ɪ/ and Mandarin /ɛ, æ/). This can be explained by a congruence and convergence of bottom-up and top-down processing. Where a listener was unable to differentiate between target vowels, the sentence context filled the gap. The word from the minimal pair that was “closest” (most closely associated with the context) in the listener’s semantic network was presumably recalled (Collins & Loftus, 1975). Because the listener’s association was the same as the key, the listener responded correctly. Listeners who could readily differentiate between target vowels would also be expected to respond correctly given the vowel presented. Here, vowel perception (bottom-up) and semantic association (top-down) cannot be disentangled. As listeners who could readily differentiate between target vowels and listeners who could not would both be expected to respond correctly, same associations offered little information about listeners’ ability to differentiate vowels. For analysis purposes, same association responses could be treated similarly to same-different biases, where biased responses (default responses to perceptual stimuli) are omitted due to being uninterpretable or uninteresting (McGuire, 2010).

Equal associations

Scores were greater than expected for *equal* associations, where the effect of association was controlled for—to the extent possible with the experiment’s self-report survey—better enabling an assessment of vowel perception in connected speech without familiarity or association affecting results. Theoretically, neither word in the minimal pair was cued by semantic association, neither word was “closer” in the participant’s semantic network (Collins & Loftus, 1975). For an assessor wishing to examine the effects of connected speech on vowel perception, this association type presents itself as most suitable. The most interesting association type for examining the interaction between top-down and bottom-up processing, however, was the *opposite* association.

Opposite associations

Opposite associations led to a stark reduction in scores for many participants. Such a result is unsurprising considering the opposite (and incorrect) word was associated with the sentential context. Consistently incorrectly responding to opposite associations suggested that either the listener could not hear the distinction between target vowels and thus relied upon context to disambiguate what would be, for the listener, a homophonous word, or that there was an imbalance in the listener's processing where top-down is dominating bottom-up (Field, 2004). Consistently responding correctly in opposite association cases suggested that the listener was able to make effective use of bottom-up processing. Despite associating the non-key word from the minimal pair more strongly with the sentence context (suggesting the non-key word would be more readily accessible to the listener's semantic network), the listener was able to respond accurately with the word expected to be more distant in the semantic network. As perception is indirectly observed and the listener's response is the only way to determine accurate perception, it may be surmised that the listener correctly perceived the listening prompt as a result of effective bottom-up processing.

The opposite associations the most difficult prompt type condition examined, and difficulty was considered relevant to the construct (Messick, 1995). First, because dialogue outside of a laboratory setting is organic and generative, listeners will find themselves in situations where they must perceive words in contexts which were not the first words that would typically come to mind. Regardless of expectation, listeners must still navigate their environment. An assessment which contains this type of prompt may identify a need (i.e., an inability to perceive vowels accurately in ambiguous situations), and this need would otherwise go undetected by isolated word prompt types or connected speech prompts which have been controlled for association. Second, doing well on opposite association sentences entailed doing well in all other conditions (and prompts). This can be viewed as the opposite end of the spectrum from bVt, where doing poorly on bVt typically translated to doing poorly elsewhere. It discriminates strongly. Consequently, this would be helpful because fewer items may be required for an assessment, saving time for additional targets or assessments if needed. Finally, Diverse Sentences opposite is the only prompt type which unequivocally engages both top-down and bottom-up processes. A listener must use all available information to understand speech, so it may also help identify a processing imbalance, where bottom-up needs to be strengthened to facilitate effective perception.

Having discussed the generalisability of bVt prompts, the potential suitability of employing more diverse prompt types for assessing L2 vowel perception, and the effect of association on L2 vowel perception in sentences, the thesis now concludes.

Conclusion

This research examined the use of diverse prompt types to address a scarcity of published literature investigating L2 vowel perception beyond isolated, fixed consonantal frames. The diverse prompt types were introduced to (1) uncover the extent to which a fixed, isolated frame (bVt) generalises to phonologically diverse and sentential contexts, and (2) to uncover the suitability of diverse prompt types for empirical inquiry. Two studies investigated the effects of using bVt and more diverse prompt types for assessing intermediate-advanced adult L2 perception of English /i/-/ɪ/ and /ɛ/-/æ/ vowel pairs, framing results within Best and Tyler’s (2007) Perceptual Assimilation Model-L2 (PAM-L2). The studies measured internal consistency, congruence with PAM-L2 predictions, and listeners’ subjective experiences with each prompt type. Generalised linear mixed model analyses served to explore how well listener performance with canonical bVt prompts predicted performance with the more diverse prompt types and to uncover the relative effects of sentence-specific variables for perceiving L2 vowels.

The generalisability of bVt to connected speech was found to be negligible. Diverse Sentence prompts were found to better match PAM-L2 predictions than isolated prompts and held sufficiently strong internal consistency, which is consistent with standards for empirical inquiry. Participant feedback displayed benefits and detriments for both isolated and connected speech prompts. Sentential context was viewed by some participants as helpful due to factors such as context and increased attention, while for at least one participant the context was distracting. The fixed, isolated prompt type permitted participants to focus on a specific vowel, but was seen as repetitive and potentially compromised listeners' attention. More work is needed to explore the link between prompt type, attention, and performance. The relationship between target words and their context was examined, offering insight into the interaction between bottom-up and top-down processing for vowel perception. A potential hierarchy was found, where strong performance with the "opposite" associations indicated strong performance with bVt, but strong performance with bVt did not indicate strong performance with "opposite" associations. Given the limited generalisability of bVt, the relatively strong performance of the diverse, connected speech prompts, and the potential for connected speech prompts to provide additional information about listeners, use of diverse sentences may be considered a viable complement or replacement for isolated, fixed consonantal frames for assessing vowel perception.

Contributions and limitations

The present study offers several contributions to the literature. First, it re-examines classically employed stimuli and uncovers a systematic, implicit leaning toward construct underrepresentation over construct irrelevant variance. This is displayed through the canonical use of fixed, isolated words (and syllables) over connected speech in the absence of published empirical support which illustrates connected speech prompts are problematic. The conventionally employed, fixed consonantal frame prompt conceptually underrepresented the construct of vowel perception—focusing strictly on bottom-up processes and an unnaturally restricted phonological context—did not generalise well beyond its context. The study offers novel stimuli for exploration and introduces analyses to account for features inherent in the new stimuli (e.g., mixed effects modelling). Congruent with contemporary psychological assessment practices, item analysis was employed in a pilot to identify well-performing stimuli and eliminate poorly-performing stimuli.

Contra to predominant vowel perception materials development, the study suggests that variations in phonological environments and the consequent relative difficulty of each stimulus as an important feature for vowel perception assessment development. By incorporating phonological variation and expanding isolated words to connected speech, the study helps provide a foundational step toward a more authentic (for both materials and cognitive processes) assessment of a participant's ability to perceive the difference between systematically conflated speech sounds. Finally, the study addresses an underexamined component of the contemporary intelligibility paradigm, where comprehensibility may be high where intelligibility is low. This was identified as a non-trivial concern where accuracy and intelligibility intersect.

Contributions herein provide potential grounds for both prospective and retrospective inquiry. Revisiting seminal works with the original prompts complemented with more diverse, connected speech prompt types may yield a deeper, possibly modified understanding of L2 vowel perception.

When designing this research, scope concessions were made for practicality, narrowing the focus to front vowels. Despite the limited scope, however, this work conceptually supports an approach to prompt design that may generalise to speech perception assessments in general, regardless of segment, suprasegment, or language. Specifically, for exploring an individual's ability to differentiate between contrastive L2 features, this research suggests utility in developing prompts which better reflect the construct than an isolated prompt type. Additional research (post-doctoral studies, replication studies) will be needed to help strengthen the research findings and fill theoretical gaps that may emerge.

Phonetic context is potentially a confounding factor in some of your conclusions because it wasn't controlled in the diverse sentences condition. Further work may be done to examine effects of phonological and phonetic contexts and how those effects compare or interact with predictors (e.g., opposite associations) explored in this study.

Though this research incorporated participant feedback, it may only be considered preliminary. A more probing survey or interview is recommended in future research to further explore interactivity between the participant and the assessment. This research maximised emphasis on prompts, both in type and number. The large number of items presented in the two studies prevented a deeper, more targeted exploration into the participant experience. Given the internal consistency displayed with a reduced number of items (as found with the adjusted Cronbach's alpha), the number of items presented may be substantially reduced, enabling greater opportunity to probe participant experience without increasing time allotment.

Implications and future research

This work should be viewed as an ongoing effort to ensure prompts used for assessing L2 vowel perception adequately reflect the construct. While the sentence prompts employed herein may more fully mirror the construct of L2 listening perception than an isolated prompt type such as bVt, the domain of language use addressed by these sentential prompts remains limited. Systematically manipulating variables such as sentence type (e.g., simple, compound, complex, declarative, interrogative, etc.) and sentence length could provide a broader representation of the construct and contexts in which it is used.

Associations offered a promising glimpse into the interaction between top-down and bottom-up processes, and further work may be done to identify associations between these processes and assimilation types. For instance, PAM-L2 predicts two category discrimination will be the same as L1 discrimination, yet that was not found to not necessarily be the case in Experiment 2 where Spanish L1s performed considerably worse than expected for the /ɛ, æ/ vowel pair, particularly where associations were "opposite". There may be within- and between-assimilation-type gradations that emerge from incorporating prompts which explore these processes.

As discussed in Experiment 1 and 2, there was evidence to suggest that including target words in sentences may retain participants' attention better than isolated words when assessing L2 vowel perception. This was an intriguing, but speculative prospect, requiring greater rigor than could be provided by the present exploratory research. If supported by additional empirical study, increased attention would constitute further evidence for the utility in employing prompts which contain connected speech.

Lastly, this work has potential implication for high variability phonetic training. Though results from the present study would suggest that training on individual, isolated contexts should not be expected to generalise to diverse sentential prompts, it is possible that training using such prompts might. Thomson (2012) noted that unbounded variability would be detrimental to perception, but in the context of connected speech prompts, how much variability and for which target group (e.g., a given language, age, or proficiency) constitute relevant extensions to investigate. It may be that for some groups or contexts, the added variability yields better training results than isolated speech prompts.

References

Antonijevic, S., Durham, R., & Chonghaile, Í. N. (2017). Language performance of sequential bilinguals on an Irish and English sentence repetition task. *Linguistic Approaches to Bilingualism, 7*(3-4), 359-393.

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

Appezzato, Hackerott, M. M. S., & Avila, C. R. B. de. (2018). Speech perception task with pseudowords. *CoDAS (São Paulo), 30*(2).

Archila-Suerte, P., Bunta, F., & Hernandez, A. E. (2016). Speech sound learning depends on individuals' ability, not just experience. *International Journal of Bilingualism*, *20*(3), 231-253.

Armon-Lotem, de Jong, J., & Meir, N. (2015). *Assessing multilingual children: disentangling bilingualism from language impairment* (Vol. 13). NBN International.

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*(1), 92.

Baart. (2010). *A field manual of acoustic phonetics / Joan L.G. Baart*. Sil International.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Balas, A. (2018). English vowel perception by Polish advanced learners of English. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique, 63*(3), 309-338.

Barreda, S. (2017). An investigation of the systematic use of spectral information in the determination of apparent-talker height. *The Journal of the Acoustical Society of America, 141*(6), 4781-4792.

Barrios, S., & Hayes-Harb, R. (2021). L2 processing of words containing English/æ/-/ɛ/and/l/-/ɹ/contrasts, and the uses and limits of the auditory lexical decision task for understanding the locus of difficulty. *Frontiers in Communication, 6*, 144.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Best, C. T. (1995). A direct realist view of cross-language speech perception. *Speech Perception and Linguistic Experience*, 171-206.

Best, C. C., & McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech, 46*(2-3), 183-216.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America, 109*(2), 775-794.

Bilger, R. C. (1984). Speech recognition test development. In E. Elkins (Ed.), *Speech recognition by the hearing impaired*. ASHA Reports, 14, 2-7.

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second language competence. In E Tarone, S. Gass, & A. Cohen (Eds.), *Research Methodology in Second-Language Acquisition*, 245–261.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Bohn, O. S. (2017). Cross-Language and Second Language Speech Perception. *The Handbook of Psycholinguistics*, 213-239.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127-135.

Bogert, B. P. (1953). On the band width of vowel formants. *The Journal of the Acoustical Society of America, 25*(4), 791-792.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Braun, V. & Clarke, V. (2012) Thematic analysis. In Cooper, H. (Ed.), *The Handbook of Research Methods in Psychology*. Washington, DC: American Psychological Association.

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological Review, 74*(1), 1.

Brosseau-Lapré, F., Rvachew, S., Clayards, M., & Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics*, *34*(3), 419-441.

Brown, A. (1988). Functional load and the teaching of pronunciation. *Tesol Quarterly*, *22*(4), 593-606.

Brekelmans, G., Evans, B. G., & Wonnacott, E. (2020). *Training child learners on non-native vowel contrasts: the role of talker variability*. [Manuscript submitted for publication]. Speech Hearing and Phonetic Sciences, University College London.

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, *33*(3), 433-461.

Campbell, G. A. (1910). XII. Telephonic intelligibility. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 19*(109), 152-159.

Canale, M. (1987). The measurement of communicative competence. *Annual review of applied linguistics, 8*, 67-84.

Canfield, D. L. (1981). *Spanish pronunciation in the Americas*. University of Chicago Press.

Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(5), 629-647.

Catford, J. C. (1987). Phonetics and the teaching of pronunciation: A systemic description of English phonology. *Current Perspectives on Pronunciation: Practices Anchored in Theory*, 87-100.

Chapelle, C. A. (2020). An introduction to Language Testing's first Virtual Special Issue: Investigating consequences of language test use. *Language Testing, 37*(4), 638-645.

Chen, S. Y., Griffin, B. M., Mancuso, D., Shiau, S., DiMattia, M., Cellum, I., ... & Lalwani, A. K. (2018). The development and validation of the speech quality instrument. *The Laryngoscope, 128*(7), 1622-1627.

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods, 18*(2), 207-230.

Clarke, V., & Braun, V. (2018). Using thematic analysis in counselling and psychotherapy research: A critical reflection. *Counselling and Psychotherapy Research, 18*(2), 107-110.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English*. The Journal of the Acoustical Society of America, 116*(6), 3647-3658.

Clopper, C. G. (2021). Perception of dialect variation. *The Handbook of Speech Perception*, 333-364.

Cobb, T. (n.d.). The compleat lexical tutor. Retrieved from http://www.lextutor.ca/ on 20 August 2021.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98.

Crandall, I. B. (1925). The sounds of speech. *The Bell System Technical Journal, 4*(4), 586-639.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281.

Croot, K., Lalas, G., Biedermann, B., Rastle, K., Jones, K., & Cholin, J. (2017). Syllable frequency effects in immediate but not delayed syllable naming in English. *Language, Cognition and Neuroscience, 32*(9), 1119-1132.

Darcy, I., Dekydtspotter, L., Sprouse, R. A., Glover, J., Kaden, C., McGuire, M., & Scott, J. H. (2012). Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English–L2 French acquisition. *Second Language Research, 28*(1), 5-40.

De Deyne, S., Kenett, Y. N., Anaki, D., Faust, M., & Navarro, D. (2017). Large-scale network representations of semantics in the mental lexicon. In M. N. Jones (Ed.), *Big data in cognitive science* (pp. 174–202). Routledge/Taylor & Francis Group.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology, 30*(1).

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1-16.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *Tesol Quarterly*, *39*(3), 379-397.

Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E., & Fleming, J. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, *64*(3), 526-548.

Derwing, T. M., Rossiter, M. J., & Munro, M. J. (2002). Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingual and Multicultural Development, 23*(4), 245-259.

Dickerson, W. (2016). A practitioner’s guide to English rhythm: A return to confidence. In J. Levis, H. Le, I. Lucic, E. Simpson, & S. Vo (Eds.), *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference*, October 2015 (pp. 39-50).

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology, 58*(1), 116.

Evans, B. G., & Alshangiti, W. (2018). The perception and production of British English vowels and consonants by Arabic learners of English. *Journal of Phonetics, 68*, 15-31.

Everington, C., Notario-Smull, H., & Horton, M. L. (2007). Can defendants with mental retardation successfully fake their performance on a test of competence to stand trial?. *Behavioral Sciences & the Law, 25*(4), 545-560.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80-92.

Field, J. (1999). Bottom-up and top-down. *ELT Journal, 53*, 338–339.

Field, J. (2004). An insight into listeners' problems: Too much bottom-up or too much top-down? *System, 32*(3), 363-377.

Field, J. (2008). Bricks or mortar: which parts of the input does a second language listener rely on? *TESOL Quarterly, 42*(3), 411-432.

Field, J. (2019). *Second language listening: Current ideas, current issues*. Cambridge University Press.

Flege, J. E. (1991). Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. *The Journal of the Acoustical Society of America, 89*(1), 395-411.

Flege, J. E. (1995a). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.) *Speech perception and linguistic experience: Issues in cross-language research, 92*, 233-277.

Flege, J. E. (1995b). *Oddity discrimination of English vowels by non-natives*. Unpublished manuscript, Department of Biocommunication, University of Alabama at Birmingham.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*(4), 437-470.

Flege, J. E., & Bohn, O. S. (2021). The revised speech learning model (SLM-r). *Second language speech learning: Theoretical and empirical progress*, 3-83.

Flege, J. E., & MacKay, I. R. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition*, *26*(1), 1-34.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1996). Factors affecting the production of word-initial consonants in a second language. In R. Bayley & D. R. Preston (Eds.), *Second language acquisition and linguistic variation. Amsterdam: John Benjamins. 47–73.*

Flege, J. E., & Wayland, R. (2019). The role of input in native Spanish Late learners’ production and perception of English phonetic segments. *Journal of second language studies, 2*(1), 1-44.

Fletcher, H. (1922). The nature of speech and its interpretation. *The Bell System Technical Journal, 1*(1), 129-144.

Fletcher, H. (1929). *Speech and Hearing* (Van Nostrand, New York).

Fletcher, H., & Steinberg, J. C. (1929). *Articulation testing methods. The Bell System Technical Journal, 8*(4), 806-854.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132.

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning, 66*(2), 419-447.

Geer, J. G. (1991). Do open-ended questions measure "salient" issues?. *Public Opinion Quarterly*, *55*(3), 360-370.

Gerrits, E. (2001). *The categorisation of speech sounds by adults and children*. Netherlands Graduate School of Linguistics.

Gerrits, E., & Schouten, B. (1998). Categorical perception of vowels. In *Fifth International Conference on Spoken Language Processing*.

Gilmore, A. (2011). "I prefer not text": Developing Japanese learners' communicative competence with authentic materials. *Language learning, 61*(3), 786-819.

Gokgoz-Kurt, B., & Holt, D. E. (2018). Connected Speech in Advanced-Level Phonology. *The Handbook of Advanced Proficiency in Second Language Acquisition*, 304-322.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28*(5), 501-518.

Grant, K. W., & Seitz, P. F. (2000). The recognition of isolated words and words in sentences: Individual variability in the use of sentence context. *The Journal of the Acoustical Society of America, 107*(2), 1000-1011.

Grier, R. A. (2015, September). How high is high? A meta-analysis of NASA-TLX global workload scores. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 59, No. 1, pp. 1727-1731). Sage CA: Los Angeles, CA: SAGE Publications.

Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly*, *26*(2), 385-390.

Guariento, W., & Morley, J. (2001). Text and task authenticity in the EFL classroom. *ELT journal, 5*5(4), 347-353.

Haertel, E. H. (2006). Reliability. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger.

Halberstadt, J. B., Niedenthal, P. M., & Kushner, J. (1995). Resolution of lexical ambiguity by emotional state. *Psychological Science, 6*(5), 278-282.

Harding, L. (2017). Validity in pronunciation assessment. In *Assessment in second language pronunciation* (pp. 42-60). Routledge.

Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage publications.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology* (Vol. 52, pp. 139-183). North-Holland.

Hazan, V., & Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: speaker and listener effects. *Language and Speech, 43*(3), 273-294.

Hillenbrand, J. M. (2013). Static and dynamic approaches to vowel perception. In *Vowel inherent spectral change* (pp. 9-30). Springer, Berlin, Heidelberg.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America, 97*(5), 3099-3111.

Hirata, Y., Whitehurst, E., & Cullings, E. (2007). Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. *The Journal of the Acoustical Society of America*, *121*(6), 3837-3845.

Hoekstra, R., Vugteveen, J., Warrens, M. J., & Kruyen, P. M. (2019). An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology, 22*(4), 351-364.

Højen, A., & Flege, J. E. (2006). Early learners' discrimination of second-language vowels. *The Journal of the Acoustical Society of America, 119*(5), 3072-3084.

Holt, L., & Wade, T. (2004). Non-linguistic sentence-length precursors affect speech perception: Implications for speaker and rate normalization. In *Proceedings from sound to sense: Fifty+ years of discoveries in speech communication*.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America, 25*(1), 105-113.

Huang, B., & Liao, X. (1997). Modern Chinese (in Chinese). *The Beijing High Education Press, 127*, 126-127.

Inceoglu. (2022). Language Experience and Subjective Word Familiarity on the Multimodal Perception of Non-native Vowels. *Language and Speech, 65*(1), 173–192.

Ingram, J. C., & Park, S. G. (1997). Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics, 25*(3), 343-370.

Ingvalson, E. M., Ettlinger, M., & Wong, P. C. (2014). Bilingual speech perception and learning: A review of recent trends. *International Journal of Bilingualism*, *18*(1), 35-47.

Isaacs, T. (2016). Assessing speaking. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 131–146). DeGruyter Mouton.

Isaacs, T., & Harding, L. (2017). Pronunciation assessment. *Language Teaching*, *50*(3), 347-366.

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America, 126*(2), 866-877.

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145-160.

Janssen, D. P. (2012). Twice random, once mixed: Applying mixed models to simultaneously analyze random effects of language and participants. *Behavior Research Methods, 44*(1), 232-247.

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America, 119*(2), 1118-1130.

Jones, J. (2015). *Exploring open consonantal environments for testing vowel perception*. Unpublished Master's thesis, University of Melbourne, Melbourne, Australia.

Jones, J., & Isaacs, T. (2022). Assessing second language pronunciation. In H. Moehebbi, C. Coombe (Eds.), *Research questions in language education: A reference guide for teachers* (pp. 305-310). Springer.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198-211.

Kang, O., Thomson, R. I., & Moran, M. (2020). Which features of accent affect understanding? Exploring the intelligibility threshold of diverse accent varieties. *Applied Linguistics, 41*(4), 453-480.

Katz, J. (2012). Compression effects in English. *Journal of Phonetics, 40*(3), 390-402.

Khan, I. A. (2016). Difficulties in Mastering and Using English for Specific Purpose (Medical Vocabulary): A Linguistic Analysis of Working Saudi Hospital Professionals. *International Journal of Education*, *8*(1), 78-93.

Kim, J. E. (2010). Perception and production of English front vowels by Korean speakers. *Phonetics and Speech Sciences, 2*(1), 51-58.

Kim, D., Clayards, M., & Goad, H. (2018). A longitudinal study of individual differences in the acquisition of new vowel contrasts. *Journal of Phonetics, 67*, 1-20.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics, 3*(3), 129-140.

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics, 63*(8), 1421-1455.

Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing*, 0265532217710049.

Kondaurova, M. V., & Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *The Journal of the Acoustical Society of America, 124*(6), 3959-3971.

Kuhl, P. K. (1992). Psychoacoustics and speech perception: internal standards, perceptual anchors, and prototypes. In L. A. Werner & E. W.Rubel (Eds.) *Developmental psychoacoustics*, pp. 293–332. American Psychological Association.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 979-1000.

Ladefoged, P., & Disner, S. F. (2012). *Vowels and consonants*. John Wiley & Sons.

Ladefoged, P., & Johnson, K. (2014). *A course in phonetics*. Cengage learning.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Lee, H. Y., Hwang, H., & Yoon, E. (2016). Effects of perceptual training on the perception of Korean boundary tones by Chinese learners. *The Journal of the Acoustical Society of America*, *140*(4), 3341-3341.

Lee, H., & Jongman, A. (2016). A diachronic investigation of the vowels and fricatives in Korean: An acoustic comparison of the Seoul and South Kyungsang dialects. *Journal of the International Phonetic Association, 46*(2), 157-184.

Lee, S., & Cho, M. H. (2020). The impact of L2-learning experience and target dialect on predicting English vowel identification using Korean vowel categories. *Journal of Phonetics, 82*, 100983.

Lee, S. W., Lee, H. W., & Kim, C. G. (2015). A System Design for Audio Level Measurement based on ITU-R BS. 1770-3. In *Information Science and Applications* (pp. 1101-1105). Springer.

Lee, O. J., & Xiong, Y. (2021). Distribution of the Mandarin vowels in typological perspective. *Linguistic Research, 38*(2), 329-363.

Lehiste, I., & Peterson, G. E. (1961). Transitions, glides, and diphthongs. *The journal of the acoustical society of America, 33*(3), 268-277.

Lengeris, A. (2009). Perceptual assimilation and L2 learning: Evidence from the perception of Southern British English vowels by native speakers of Greek and Japanese. *Phonetica, 66*(3), 169-187.

Leong, C. X. R., Price, J. M., Pitchford, N. J., & van Heuven, W. J. (2018). High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PloS one, 13*(10), e0204888.

Levy, E. S., & Strange, W. (2008). Perception of French vowels by American English adults with and without French language experience. *Journal of Phonetics*, *36*(1), 141-157.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English/r/and/l: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874-886.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1993). Training listeners to perceive novel phonetic categories: How do we know what is learned?. *The Journal of the Acoustical Society of America, 94*(2), 1148-1151.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software, 6*(60).

Lüdecke, D., Patil, I., Ben-Shachar, M. S., Wiernik, B. M., Waggoner, P., & Makowski, D. (2021). See: an R package for visualizing statistical models. *Journal of Open Source Software, 6*(64), 3393.

Macmillan, & Creelman, C. D. (2005). *Detection theory : a user's guide / Neil A. MacMillan and C. Douglas Creelman*. (2nd ed.). Psychology Press.

Magimairaj, B. M., Capin, P., Gillam, S. L., Vaughn, S., Roberts, G., Fall, A. M., & Gillam, R. B. (2022). Online Administration of the Test of Narrative Language–Second Edition: Psychometrics and

Considerations for Remote Assessment. *Language, Speech, and Hearing Services in Schools*, 1-13.

Maltenfort, M. G., Restrepo, C., & Chen, A. F. (2020). *Statistical Reasoning for Surgeons*. CRC Press.

Maxwell, O., Baker, B., Bundgaard-Nielsen, R., & Fletcher, J. (2015). A comparison of the acoustics of nonsense and real word stimuli: coronal stops in Bengali. In Scottish Consortium for ICPhS (Ed.) *Proceedings of the 18th International Congress of Phonetic Sciences*, University of Glasgow, United Kingdom (2015), pp. 1-5.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256.

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech, 58*(4), 502-521.

McGuire, G. (2010). A brief primer on experimental designs for speech perception research. *Laboratory Report*, *77*.

McNamara, T. (2000). *Language testing*. Oxford University Press.

Mitterer, H., & Mattys, S. L. (2017). How does cognitive load influence speech perception? An encoding hypothesis. *Attention, Perception, & Psychophysics*, *79*(1), 344-351.

Monteiro, K., & Kim, Y. (2020). The effect of input characteristics and individual differences on L2 comprehension of authentic and modified listening tasks. *System, 94*, 102336.

Morrison, G. S. (2013). Theories of vowel inherent spectral change. In *Vowel inherent spectral change* (pp. 31-47). Springer.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*(4), 520-531.

Munro, M. J., & Derwing, T. M. (2015). Intelligibility in research and practice: Teaching priorities. *The handbook of English pronunciation*, 377-396.

Nagle, C. L., & Baese-Berk, M. M. (2021). Advancing the State of the Art in L2 Speech Perception-Production Research: Revisiting Theoretical Assumptions and Methodological Practices. *Studies in Second Language Acquisition*, 1-26.

Nelson, C. L., & Kang, S. Y. (2015). Pronunciation and world Englishes. *The handbook of English pronunciation*, 320-329.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 113–134).

Nevo, B. (1980). Item analysis with small samples. *Applied Psychological Measurement, 4*(3), 323-329.

Njie, S., Lavan, N., & McGettigan, C. (2022). Talker and accent familiarity yield advantages for voice identity perception: a voice sorting study. *Memory & Cognition*, 1-13.

Nunn, T. B., Jiang, D., Green, T., Boyle, P. J., & Vickers, D. A. (2019). A systematic review of the impact of adjusting input dynamic range (IDR), electrical threshold (T) level and rate of stimulation on speech perception ability in cochlear implant users. *International Journal of Audiology, 58*(6), 317-325.

Ockey, G. J., & Wagner, E. (2018). *Assessing L2 listening: Moving towards authenticity* (Vol. 50). John Benjamins Publishing Company.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: debates and practical guidelines. *International Journal of Qualitative Methods, 19*, 1609406919899220.

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2016). Eye movements reveal fast, voice-specific priming. *Journal of Experimental Psychology: General, 145*(3), 314.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy*, 174-200. Cambridge University Press.

Park, Y. S. (2019). Item Response Theory. In *Assessment in health Professions education* (pp. 287-297). Routledge.

Pelzl, E., Lau, E. F., Guo, T., & DeKeyser, R. (2019). Advanced second language learners' perception of lexical tone contrasts. *Studies in Second Language Acquisition, 41*(1), 59-86.

Purdy, S. C., Welch, D., Giles, E., Morgan, C. L. A., Tenhagen, R., & Kuruvilla-Mathew, A. (2017). Impact of cognition and noise reduction on speech perception in adults with unilateral cochlear implants. *Cochlear Implants International, 18*(3), 162-170.

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America, 130*(1), 461-472.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*(2), 175-184.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America, 32*(6), 693-703.

Pinker, S., & Birdsong, D. (1979). Speakers' sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior, 18*(4), 497-508.

Pinner, R. (2014). The authenticity continuum: Towards a definition incorporating international voices: Why authenticity should be represented as a continuum in the EFL classroom. *English today, 30*(4), 22-27

Polka, L., & Bohn, O. S. (2011). Natural Referent Vowel (NRV) framework: An emerging view of early phonetic development. *Journal of Phonetics, 39*(4), 467-478.

Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *The Journal of the Acoustical Society of America*, *22*(6), 807-820.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language, 29*(6), 633-654.

Powell, M. J. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives*. Cambridge NA Report NA2009/06, University of Cambridge, Cambridge, 26.

Pruitt, J. S., Jenkins, J. J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *The Journal of the Acoustical Society of America*, *119*(3), 1684-1696.

Purdy, S. C., Welch, D., Giles, E., Morgan, C. L. A., Tenhagen, R., & Kuruvilla-Mathew, A. (2017). Impact of cognition and noise reduction on speech perception in adults with unilateral cochlear implants. *Cochlear Implants International, 18*(3), 162-170.

QSR International Pty Ltd. (2018) *NVivo* (Version 12), Retrieved from https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing* [Software]. Retrieved from https://www.r-project.org.

Rauber, A. S., Rato, A., Kluge, D. C., & Santos, G. (2012). *TP* (Version 3.1) [Software]. Retrieved from http://www.worken.com.br/tp_regfree.php.

Richardson, E. G. (1940). Recent work in experimental phonetics. *Nature, 145*(3683), 841.

Rogers, L. (2017). Introduction and History of Sensory Discrimination Testing. In *Discrimination Testing in Sensory Science* (pp. 3-30).

Ronquest, R. (2018). Vowels. In K. Geeslin (Ed.), *The Cambridge Handbook of Spanish Linguistics* (pp. 145-164). Cambridge University Press.

Rosenblum, L. D., & Dorsi, J. (2021). Primacy of multimodal speech perception for the brain and science. *The Handbook of Speech Perception*, 28-57.

Rusticus, S. A., & Lovato, C. Y. (2014). Impact of sample size and variability on the power and type I error rates of equivalence tests: A simulation study. *Practical Assessment, Research, and Evaluation, 19*(1), 11.

Saito, K. (2018). Advanced 15 Second Language Segmental and Suprasegmental Acquisition. *The Handbook of Advanced Proficiency in Second Language Acquisition*, 282.

Schäfer, R. (2020). Mixed-Effects Regression Modeling. In *A practical handbook of corpus linguistics* (pp. 535-561). Springer.

Schouten, B., Gerrits, E., & van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication*, *41*(1), 71-80.

Shin, J. (2015). Vowels and consonants. In L. Brown & J. Yeon (Eds.), *The handbook of Korean linguistics* (pp. 3-21). Wiley.

Sim, S. M., & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals-Academy of Medicine Singapore, 35*(2), 67.

Snowling, M. J., Lervåg, A., Nash, H. M., & Hulme, C. (2019). Longitudinal relationships between speech perception, phonological skills and reading in children at high-risk of dyslexia. *Developmental Science, 22*(1), e12723.

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech, 53*(2), 273-293.

Solso, R. L., & Juel, C. L. (1980). Positional frequency and versatility of bigrams for two-through nine-letter English words. *Behavior Research Methods & Instrumentation, 12*(3), 297-343.

Spoehr, K. T., & Smith, E. E. (1973). The role of syllables in perceptual processing. *Cognitive Psychology, 5*(1), 71-89.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly,* 32-71.

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research, 61*(6), 1463-1486.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. A., & Nishi, K. (2001). Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners. *The Journal of the Acoustical Society of America*, *109*(4), 1691-1704.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131-145.

Strange, W., Weber, A., Levy, E. S., Shafiro, V., Hisagi, M., & Nishi, K. (2007). Acoustic variability within and across German, French, and American English vowels: Phonetic context effects. *The Journal of the Acoustical Society of America*, *122*(2), 1111-1129.

Sun, L., & van Heuven, V. J. (2007). Perceptual assimilation of English vowels by Chinese listeners: Can native-language interference be predicted? *Linguistics in the Netherlands, 24*(1), 150-161.

Swain, J. (2018). *A hybrid approach to thematic analysis in qualitative research: Using a practical example*. SAGE Publications Ltd.

Taber. K. S. (2017). The Use of Cronbach’s Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296.

Tang, W., Cui, Y., & Babenko, O. (2014). Internal consistency: Do we really know what it is and how to assess it. *Journal of Psychology and Behavioral Science, 2*(2), 205-220.

Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). Thematic analysis. *The SAGE handbook of qualitative research in psychology, 2*, 17-37.

Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning*, *62*(4), 1231-1258.

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

Thomson, R. I. (2018). High variability [pronunciation] training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation, 4*(2), 208-231.

Thomson, R. I., & Derwing, T. M. (2016). Is phonemic training using nonsense or real words more effective? In J. Levis, H. Le., I. Lucic, E. Simpson, & S. Vo (Eds.). *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Dallas, TX, Oct. 2015. (pp. 88-97). Ames, IA: Iowa State University.

Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Flege, J. (2005). A developmental study of English vowel production and perception by native Korean adults and children. *Journal of Phonetics, 33*(3), 263-290.

Tyler, M. D. (2019). PAM-L2 and phonological category acquisition in the foreign language classroom. *A sound approach to language matters–In honor of Ocke-Schwen Bohn*, 607-630.

Tyler, M. D. (2021). Phonetic and phonological influences on the discrimination of non-native phones. *Second language speech learning: Theoretical and empirical progress*, 157-174.

Tyler, M. D., Best, C. T., Faber, A., & Levitt, A. G. (2014). Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica, 71*(1), 4-21.

Uddenberg, S., & Shim, W. M. (2015). Seeing the world through target-tinted glasses: Positive mood broadens perceptual tuning. *Emotion, 15*(3), 319.

Uddin, S., Reis, K. S., Heald, S. L., Van Hedger, S. C., & Nusbaum, H. C. (2020). Cortical mechanisms of talker normalization in fluent sentences. *Brain and Language, 201*, 104722.

Verbyla, A. P. (2019). A note on model selection using information criteria for general linear models estimated using REML. *Australian & New Zealand Journal of Statistics, 61*(1), 39-50.

Wagner, E. (2014). Using unscripted spoken texts in the teaching of second language listening. *TESOL Journal, 5*(2), 288-311.

Wagner, E. (2021). Assessing listening. In *The Routledge Handbook of Language Testing* (pp. 223-235). Routledge.

Wagner, E., & Toth, P. D. (2017). *The Role of Pronunciation in the Assessment of Second Language Listening Ability*. Multilingual Matters.

Wang, X., & Munro, M. J. (1999, August). The perception of English tense-lax vowel pairs by native Mandarin speakers: The effect of training on attention to temporal and spectral cues. In *Proceedings of the 14th international congress of phonetic sciences* (Vol. 3, pp. 125-128). Berkeley, CA: University of California.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281-300.

Williams, D., Escudero, P., & Gafos, A. (2018). Spectral change and duration as cues in Australian English listeners' front vowel categorization. *The Journal of the Acoustical Society of America*, *144*(3).

Marslen-Wilson, W., & Tyler, L. K. (1975). Processing structure of sentence perception. *Nature, 257*(5529), 784-786.

Winer, E. S., & Snodgrass, M. (2015). Signal Detection Theory. In M. Matthen (Ed.), *The Oxford Handbook of Philosophy of Perception*.

Xu, W., & Zammit, K. (2020). Applying thematic analysis to education: A hybrid approach to interpreting data in practitioner research. *International Journal of Qualitative Methods, 19*.

Xu, A., & Lee, A. (2018). Perception of vocal attractiveness by Mandarin native listeners. In *Proceedings of the International Conference on Speech Prosody* (pp. 344-348).

Yager, J. (2016). International Medical Graduate Physician Training in American Psychiatry: Where We Are and Where We Are Going. In *International Medical Graduate Physicians* (pp. 3-10). Springer.

Yang, M., Wang, M., & Dong, G. (2020). Bayesian variable selection for mixed effects model with shrinkage prior. *Computational Statistics, 35*(1), 227-243.

Appendices

Talker word lists

The attached document was provided to talkers. Not all words and sentences were used in the pilot study. Words and utterances that were not used in the pilot will not be used in the primary study, but may be used at a later time, such as for practice in future training applications.

Voice Actor Project: Word and Sentence Lists

Upwork Job Listing: “Male and female British voice over actors needed!”

Thanks for taking part in this project! Your voice will be used for a research project on vowel perception and will be later used in a mobile application. Please read each sentence as naturally as possible. Try to read the sentences at a typical speed, not too fast or too slow. Unless told otherwise, your sample was in the range I’m looking for. Roughly, it should be 150 words per minute. Speakers tend to read lists at different speeds when at the beginning, middle, or end, so be aware of this and fight the urge!

Pronunciation:

- I have tried to add notes where I think pronunciation might be uncertain. If you are unsure about how to say a word, please use https://en.oxforddictionaries.com.

Quick note on recording:

- If using a printed version of the lists herein, please be mindful of the sounds paper makes! If using an electronic version of the list, it is best to use a device that does not have a “hum”, like most desktops and some laptops do. Tablets and mobile phones are excellent solutions.

The word lists begin on the next page.

PART 1. WORDS.

For this part, you will see a list of words. Each word should be said three times in the carrier sentence: I said [word].

- e.g., For the word *bid*, read: I said, “bid”. I said, “bid”. I said, “bid”.

Note: don’t worry if one of the utterances is off if the other two are OK. I will select the best one.

Group A. These words all contain the vowel in the word “ship”.

bid
bids
bin
bit
bitten
chick
chipper
crick
Dickens
dim
dip
fickle
fill
filling
film me (the m of “film” should be extended to the m of “me”, so it should read, filmme rather than film.me.)
fist
fizz
gin
click
grid
hid
hill
hit
him
hip
ill
keratin (pronunciation here: https://en.oxforddictionaries.com/definition/keratin)
kipper
licking
lid
lip
litter
live
liver
mill
pick
pill
pitch
pitter
relived (stress on the second syllable, not the first. i.e., reLIVED)
rich
rip
risen
Ship Lane
sick
sill
sin
sip
skim
slick
slipping
slit
strict
tinny
Whilton (the “h” is silent)
Whitfield (the “h” is silent)
whiz (the “h” is silent)
wick
wicker basket
will
willed
win
witt
Simmons Road
scrim

Group B. These words all contain the vowel in the word “sheep”.

bead
beads
bean
beat
beaten
cheek
cheaper
creak
deacons
deem
deep
fecal
feel
feeling
feel me
feast
fees
gene
clique (pronounced: cleek)
greed
heed
heel
heat
heme
heap
eel
keeper
leaking
lead
leap
litre
leave
lever
meal
peek
peel
peach
peter
relieved
reach
reap
reason
Sheep Lane
seek
seal
scene
seep
seem

scheme
sleek
sleeping
sleet
streaked
teeny
Wheelton (the “h” is silent)
Wheatfield (the “h” is silent)
Wheeze (the “h” is silent)
week
weaker basket
wheel (the “h” is silent)
wheeled (the “h” is silent)
ween
wheat (the “h” is silent)
carotene (pronunciation here: https://en.oxforddictionaries.com/definition/carotene)
Siemens Road (pronounced: seemans road)
scream

Group C. These words all contain the vowel in the word “head”.

~~a guest~~
adept
beck
bed
~~bedpan~~
bend
bending
bet
Betty
blend
~~commend~~
~~dense~~
edit
effluence
effluent
energy
epical
essay
ester
excess
exon
expend
expensive
fellow
fest
fester
fetter
flesh
gem
guess
head
hem
kettle
~~left~~
lend
leapt
men
mend
mental
mess
petter
peck
pecked
pecking
pedal
pellet
pen
penned
pest
pet
rent
send
sender
temp
temper
trek
wreck
lemon
Ecton
Ellen
~~Kent~~
Brendon Street
Edison Road
Ellen Street
Epple Road
Fenn Street
Grenville Place
Hetley Road
Kemble Road
Redcliffe Square

Group D. These words all contain the vowel in the word “had”.

~~aghast~~
adapt
back
bad
~~bad pan~~
band
banding
bat
batty
bland
~~command~~
~~dance~~
adit
affluence
affluent
anergy
apical
assay
access
axon
expand
expansive
fallow
~~fast~~
faster
fatter
flash
jam
gas
had
ham
cattle
lamin
land
lapped
~~laughed~~
man
manned
mantle

mass
pack
packed
packing
paddle
pallet
pan
panned
~~past~~
pat
patter
rant
sand
sander
tamp
tamper
track
rack
Acton
Allen
~~Kant~~
Aster
Addison Road
Allen Street
Apple Road
Brandon Street
Campbell Road
Granville Place
Hatley road
Fann Street
Radcliffe Square

Group E. These words all contain the vowel in the word "shot".

hot
hawed (pronounced: hod)
bod
bot
rot
cot
potter

Group F. These words all contain the vowel in the word "shoot".

hoot
who'd
booed
boot
root
coot
pooter

PART 2. SENTENCES.

Please read the following sentences as naturally as possible. Like you did with words, please read these three times. I will select the best of the three.

- e.g., For number 1, you'd read, "Meet me at Ship Lane. Meet me at Ship Lane. Meet me at Ship Lane".

Group A. These words all contain the vowel in the word "ship".

1. Meet me at Ship Lane.
2. Meet me at Simmons Road.

3. Book us a room in Whitfield.
4. Book us a room in Whilton.
5. The man was bitten.
6. Many impoverished people have risen to become leaders.
7. Threats scared the Dickens out of the church.
8. The Dutch have basic mills.
9. The window needs a new sill.
10. Fill the cavity first.
11. Therapy dogs should be chipper.
12. He happens to be sick.
13. Make a fist for me.
14. The old man whizzed past me.
15. That's one of the wicker baskets.
16. Take the lid for me.
17. I think that's a kipper.
18. The embroider wants you to do his bidding.
19. You'll find countless bids on eBay.
20. The body needs keratin.
21. It's a tinny audio file.
22. It was hard to see through the slit.
23. Patients keep slipping on the floors.
24. Avoid any deadly sins.
25. Skimming helps in this profession.
26. Hit the pool for 2 laps.
27. A pill will clear your skin up.
28. I associate certain gins with happiness.
29. The elderly man doesn't want to live.
30. Don't pick at it.
31. They conducted a fickle analysis.
32. "Bit," I said.
33. I said, "bit", earlier today.

Group B. These words all contain the vowel in the word "sheep".

1. Meet me at Sheep Lane.
2. Meet me at Siemens Road.
3. Book us a room in Wheatfield.
4. Book us a room in Wheelton.
5. The man was beaten.
6. Many impoverished people have reason to become leaders.
7. Threats scared the deacons out of the church.
8. The Dutch have basic meals.
9. The window needs a new seal.
10. Feel the cavity first.
11. Therapy dogs should be cheaper.

12. He happens to be sick.
13. Make a feast for me.
14. The old man wheezed past me.
15. That’s one of the weaker baskets.
16. Take the lead for me.
17. I think that’s a keeper.
18. The embroider wants you to do his beading.
19. You’ll find countless beads on eBay.
20. The body needs carotene.
21. It’s a teeny audio file.
22. It was hard to see through the sleet.
23. Patients keep sleeping on the floors.
24. Avoid any deadly scenes.
25. Scheming helps in this profession.
26. Heat the pool for 2 laps.
27. A peel will clear your skin up.
28. I associate certain genes with happiness.
29. The elderly man doesn’t want to leave.
30. Don’t peek at it.
31. They conducted a fecal analysis.
32. “Beat,” I said.
33. I said, “beat”, earlier today.

Group C. These words all contain the vowel in the word "head".

1. Meet me at Redcliffe Square.
2. Meet me at Ellen Street.
3. Meet me at Kemble Road.
4. Meet me at Grenville Place.
5. Meet me at Fenn Street.
6. Meet me at Epple Road.
7. Meet me at Edison Road.
8. Book us a room in East Ecton.
9. I just wrecked the pool table.
10. It will be an expensive study.
11. Take a biopsy of that mess.
12. ~~A bad pest can haunt you.~~
13. I'd like to find a shop that sells gems.
14. ~~I left during the performance.~~
15. ~~You should commend new recruits.~~
16. ~~Children try to avoid bedtimes.~~
17. ~~We only have a few bedpans.~~
18. The athlete leapt everyone.
19. I'd like a pedal board for my birthday.
20. Critics penned several recent articles.
21. ~~The economist advised using PEST analysis.~~
22. Globalization brought effluence to China
23. Good coffee requires a good temper.
24. Calculate the betting averages.
25. I don't think renting is a good idea.
26. ~~I'm a guest here.~~
27. Teenagers may encounter a social enemy.
28. Locate the efferent neuron.
29. Look at the genes and exons.
30. Look for chromosome bending.
31. It can be challenging to induce energy.
32. A failed edit led to numerous problems.
33. She's a little Fleshy.
34. "Bet," I said.
35. I said, "bet", earlier today.

Group D. These words all contain the vowel in the word "had".

1. Meet me at Radcliffe Square.
2. Meet me at Allen Street.
3. Meet me at Campbell Road.
4. Meet me at Granville Place.
5. Meet me at Fann Street.
6. Meet me at Apple Road.
7. Meet me at Addison Road.
8. Book us a room in East Acton.
9. I just racked the pool table.
10. It will be an expansive study.
11. Take a biopsy of that mass.
12. ~~A bad past can haunt you.~~
13. I'd like to find a shop that sells jams.
14. ~~I laughed during the performance.~~
15. ~~You should command new recruits.~~
16. ~~Children try to avoid bad times.~~
17. ~~We only have a few bad pans.~~
18. The athlete lapped everyone.
19. I'd like a paddle board for my birthday.
20. Critics panned several recent articles.
21. ~~The economist advised using past analysis.~~
22. Globalization brought affluence to China
23. Good coffee requires a good tamper.
24. Calculate the batting averages.
25. I don't think ranting is a good idea.
26. ~~I'm aghast here.~~
27. Teenagers may encounter a social anomie.
28. Locate the afferent neuron.
29. Look at the genes and axons.
30. Look for chromosome banding.
31. It can be challenging to induce anergy.
32. A failed adit led to numerous problems.
33. She's a little flashy.
34. "Bat," I said.
35. I said, "bat", earlier today.

Group E. These words all contain the vowel in the word “hot”.

“Hot”, I said.
“Hawed”, I said.
“Bod”, I said.
“Bot”, I said.
“Rot”, I said.
“Cot”, I said.
“Potter”, I said.

Group F. These words all contain the vowel in the word “shoot”.

“Hoot”, I said.
“Who’d”, I said.
“Booed”, I said.
“Boot”, I said.
“Root”, I said.
“Coot”, I said.
“Pooter”, I said.

Listener information sheet

Participant Information Sheet For Adult Listeners

UCL Research Ethics Committee Approval ID Number: Z6364106/2018/04/149

YOU WILL BE GIVEN A COPY OF THIS INFORMATION SHEET

Title of Study: Testing canonical stimuli in speech perception research
Department: Culture, Communication and Media
Name and Contact Details of the Researcher(s): Johnathan Jones, [redacted]
Name and Contact Details of the Principal Researcher: Dr. Talia Isaacs, [redacted]

1. **Invitation Paragraph**
   My name is Johnathan Jones and I am inviting you to take in part in my voluntary study on accurate perception of English vowels. I am a post-graduate research student at the UCL Institute of Education's Department of Culture, Communication and Media. Before you decide to take part in this study, it is important for you to read the following information carefully. You may discuss it with others if you wish. Please ask if there is anything you do not understand about why the research is being done and what your participation will involve.

2. **What is the project's purpose?**
   Broadly, this experiment aims to answer the following two questions:
   a. What is the effect of using diverse words and sentences for testing vowel perception in speakers of English as a second language?
   b. How do tests that use more varied words and sentences to test vowel perception compare with traditional methods that only use a small set of words?

   Results will be used to develop a mobile application that will be used for testing and training purposes. It is possible that the ensuing testing and training application will be commercially produced (i.e., paid) in the future. Participants in the experimental phases of the study will be provided with free access to the training modules of the mobile application for the vowels they were tested with.

3. **Why have I been chosen?**
   Listening participants are central to this research. It is expected that your scores will help us learn more about you, your language group, and effective testing practices. Word of mouth is our strongest recruitment tool, so if you feel this study will benefit you or others you know, we encourage you to invite your friends to take part. We hope to recruit at least 30 Mandarin, 30 Korean, 30 Japanese, and 10 English native speakers.

   Inclusion criteria:
   - Adult (18-45 years of age)
   - native speaker of Mandarin, Korean, Japanese, or UK English
   - Minimum IELTS score of 6.0 (or equivalent); n/a for native UK English speakers

   Exclusion criteria:
   - non-normal hearing ability

4. **Do I have to take part?**
   It is your choice whether or not to take part. If you do decide to take part you will be given this information sheet to keep and will be asked to sign a consent form. You can withdraw at any

time without giving a reason. If you decide to withdraw you will be asked what you wish to happen to the data you have provided up that point.

5. **What will happen to me if I take part?**
The experiment will be conducted on a single day and is expected to take approximately 45 minutes, though it may take longer depending on voluntary breaks. Before beginning the experiment, you will be asked to complete a questionnaire that will identify your history of English language exposure and demographic data. The experiment will take place at UCL's Speech Hearing and Phonetic Sciences Department, located at Chandler House, 2 Wakefield Street, London, WC1N 1PF. Once we assign you a computer to sit at and provide your headphones, your task will be to listen to a series of audio recordings (words and sentences), and select which word or sentence is different than the others. The experiment contains blocks of questions and you will have the option to take a short break after each block. At the end of the experiment, you will be asked to complete a questionnaire about your experience. We ask that you complete all sections of the experiment; however, participation is voluntary, and you may stop at any point. To opt out prior to the experiment, please contact me at the email address or phone number listed at the bottom of this page.

6. **What are the possible disadvantages and risks of taking part?**

This experiment involves listening to audio files through headphones, and to prevent any discomfort, you will have the ability to adjust the volume to a comfortable level before and during the experiment.

7. **What are the possible benefits of taking part?**
While there are no financial benefits for those people participating in the project, it is hoped that you will find the experience rewarding. You will be able to explore a typically under-examined area of language development and, should the test be successful, it may highlight an area of strength or need for improvement. Your participation will help provide evidence supporting revised testing methods, or will help support traditional approaches to speech perception testing. While the intended benefit is to provide insight into specific conditions where you may do relatively well or poorly in discriminating between specific English vowel pairs, this is an experiment and systematic tendencies may not be uncovered. Participants in the experimental phase of the study may also have the opportunity to take part in training to improve their perceptual accuracy. Please notify Johnathan if this is something that interests you.

8. **What if something goes wrong?**
If you wish to raise a complaint, you may contact the Primary Supervisor, Dr. Talia Isaacs at [redacted]. Should you feel your complaint has not been handled to your satisfaction, you may contact the Chair of the UCL Research Ethics Committee at ethics@ucl.ac.uk.

9. **Will my taking part in this project be kept confidential?**
Your participation is strictly confidential. Your name will not be recorded and your information will be anonymised.

10. **Limits to confidentiality**
    - Please note that assurances on confidentiality will be strictly adhered to unless evidence of wrongdoing or potential harm is uncovered. In such cases the University may be obliged to contact relevant statutory bodies/agencies.
    - Confidentiality will be respected subject to legal constraints and professional guidelines.

11. What will happen to the results of the research project?

The project is planned for completion in September 2020. Upon successful completion of the study, the research will be submitted as a thesis, may be published in journal or book form, and may be presented at academic conferences. The completed work will be available by electronic copy at http://libguides.ioe.ac.uk/thesesdissertations, or by hard copy on Level 5 (Dissertations) of the UCL IOE Library. Upon completion of the project, data will be stored in UCL's Data Safe Haven.

12. Data Protection Privacy Notice

Notice:

The data controller for this project will be University College London (UCL). The UCL Data Protection Office provides oversight of UCL activities involving the processing of personal data, and can be contacted at data-protection@ucl.ac.uk. UCL's Data Protection Officer is Lee Shailer and he can also be contacted at data-protection@ucl.ac.uk**.**

Your personal data will be processed so long as it is required for the research project. We will not record your name and will anonymise or pseudonymise any personal data you provide, and will endeavour to minimise the processing of personal data wherever possible.

If you are concerned about how your personal data is being processed, please contact UCL in the first instance at data-protection@ucl.ac.uk. If you remain unsatisfied, you may wish to contact the Information Commissioner's Office (ICO). Contact details, and details of data subject rights, are available on the ICO website at: https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/individuals-rights/

16. Contact for further information

Researcher	**Primary Supervisor**	**Secondary Supervisor**
Johnathan Jones	Dr. Talia Isaacs	Professor Valerie Hazan
[redacted]	[redacted]	[redacted]

Thank you for reading this information sheet and for considering taking part in this research study.

Listener consent form

A reliable pest or a reliable past? Testing canonical stimuli in speech perception research

Consent Form

To participate in this study, please complete this consent form and return to Johnathan Jones in person or at the address below.

	Yes	No
I have read and understood the information leaflet about the research.	□	□
I understand that my participation will involve a background questionnaire, a speech perception test, and a post-test questionnaire, and I agree that the researcher may use the results as described here and in the Information Sheet.	□	□
I understand that if any of my words are used in reports or presentations they will not be attributed to me.	□	□
I understand that I can withdraw from the project at any time, and that if I choose to do this, any data I have contributed will not be used.	□	□
I understand that I can contact Johnathan Jones at any time and request for my data to be removed from the project database.	□	□
I understand that the results will be reported as part of a thesis or dissertation, may be presented at academic conferences, and may be subject to publication in journal, book, or other scholarly format. This information will not be traceable to me as an individual participant.	□	□
I agree for the data I provide to be archived at the UCL Data Safe Haven after the ensuing PhD thesis has been submitted.	□	□

Name ________________________Signed ________________________

Date _____________________

Johnathan Jones
UCL Institute of Education
20 Bedford Way London WC1H 0AL
[redacted]

Language background questionnaire

Background & Language Contact Questionnaire

The following information is important for the study and is the only personal information about you that we will keep. Your personal information is confidential and will not be shared with any other groups or individuals. Please complete each part of the questionnaire.

1. Your age:____ years old.

2. Country of birth:___________________________________

3. How old were you when you came to the UK? _____ years old.

4. Your gender: □ **Female** □ **Male** □ **Prefer not to say**

5. What do you consider your native language(s) to be? (e.g., Japanese, Mandarin, Korean, etc.)

____

6. Do you have problems with your hearing? □ **YES** □ **NO**

If **YES**, please explain.

____

____

7. Have you ever taken a language test for university or residency requirements? □ **YES** □ **NO**

If **YES**, please state the test and your overall score.

____

8. At what age were you first exposed to English? ____

9. How much do you speak **English** in the following places or situations.

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
At home											
Visiting family											
With friends											
At work											

10. Approximately how many hours per week do you spend listening to English media (music, television, radio, video games, etc.)?

____

11. a. Where did your education in English take place? (check all that apply)

b. How long were you there? (e.g., 1 year)

Country	China	Japan	Korea	UK	Other (please specify)
	□	□	□	□	
Duration					

Write here if more space is needed: ____

____

12. Were your teachers native speakers of English? **Y**□ **NO** □

If **YES**, what English-speaking country(s) were they from? ________________________

____

13. Did you ever study at an international school in your home country? **YE**□ **NO** □

If **YES**, for how long? ____

14. Have you studied phonetics or phonology in university/college? **YE**□ **NO** □

If **YES**, for how long? ____

THANK YOU!!

You're now ready for the test.

Pilot

Pilot item lists

List of /i, ɪ/ prompts used as items

	Prompt	
Prompt type	/i/	/ɪ/
bVd	bead	bid
Diverse Words	beads	bids
	beat	bit
	been	bin
	carotene	keratin
	cheek	chick
	cheaper	chipper
	clique	click
	deacons	dickens
	deem	dim
	deep	dip
	eel	ill
	feast	fist
	faecal	fickle
	feel	fill
	feeling	filling
	fees	fizz
	genes	gins
	greed	grid
	heap	hip
	heat	hit
	heed	hid
	heel	hill
	heme	hymn
	keeper	kipper
	lead	lid
	leaking	licking
	leap	lip
	leave	live
	lever	liver
	litre	litter
	meal	mill
	peach	pitch
	peak	pick
	peel	pill
	peter	pitter
	reach	rich
	reap	rip
	reason	risen
	scene	sin
	scheme	skim
	scream	scrim

	seek	sick
	seal	sill
	seep	sip
	Sheep Lane	Ship Lane
	Siemens Road	Simmons Road
	sleek	slick
	sleeping	slipping
	sleet	slit
	teeny	tinny
	weak	wick
	ween	win
	wheat	wit
	Wheatfield	Whitfield
	wheel	will
	Wheelton	Whilton
	wheeze	whiz
Sentences	The embroider wants you to do his beading	The embroider wants you to do his bidding
	You'll find countless beads on eBay	You'll find countless bids on eBay
	The man was beaten	The man was bitten
	Therapy dogs should be cheaper	Therapy dogs should be chipper
	The body needs carotene	The body needs keratin
	Threats scared the deacons out of the church	Threats scared the dickens out of the church
	Make a feast for me	Make a fist for me
	They conducted a faecal analysis	They conducted a fickle analysis
	Feel the cavity first	Fill the cavity first
	I associate certain genes with happiness	I associate certain gins with happiness
	Heat the pool for 2 laps	Hit the pool for 2 laps
	I think that's a keeper	I think that's a kipper
	Take the lead for me	Take the lid for me
	The elderly man doesn't want to leave	The elderly man doesn't want to leave
	The Dutch have basic meals	The Dutch have basic meals
	Don't peek at it	Don't pick at it
	A peel will clear your skin up	A pill will clear your skin up
	Many impoverished people have reason to become leaders	Many impoverished people have risen to become leaders
	Avoid any deadly scenes	Avoid any deadly sins
	Scheming helps in this profession	Skimming helps in this profession
	The window needs a new seal	The window needs a new sill
	Meet me at Sheep Lane	Meet me at Ship Lane
	Meet me at Siemens Road	Meet me at Simmons Road

	Patients keep sleeping on the floors	Patients keep slipping on the floors
	It was hard to see through the sleet	It was hard to see through the slit

List of /ɛ, æ/ prompts used as items

	Prompt	
Prompt type	/ɛ/	/æ/
bVd	bed	bad
Diverse Words	Ecton	Acton
	adept	adapt
	Edison Road	Addison Road
	effluence	affluence
	effluent	affluent
	Ellen	Allen
	Epple Road	Apple Road
	essay	assay
	ester	aster
	exon	axon
	beck	back
	bend	band
	bending	banding
	bet	bat
	betty	batty
	blend	bland
	Brendon Street	Brandon Street
	Kemble Road	Campbell Road
	kettle	cattle
	expend	expand
	expensive	expansive
	Fenn Street	Fann Street
	fetter	fatter
	flesh	flash
	guess	gas
	gem	jam
	Grenville Place	Granville Place
	head	had
	hem	ham
	Hetley Road	Hatley Road
	Kent	Kant
	lend	land
	leapt	lapped
	men	man
	mend	manned
	mess	mass
	peck	pack
	pecked	packed
	pecking	packing

	pedal	paddle
	pellet	pallet
	pen	pan
	penned	panned
	pet	pat
	petter	patter
	wreck	rack
	Redcliffe Square	Radcliffe Square
	rent	rant
	send	sand
	sender	sander
	temp	tamp
	temper	tamper
	trek	track
Sentences	Book us a room in East Ecton	Book us a room in East Acton
	Meet me at Edison Road	Meet me at Addison Road
	A failed edit led to numerous problems.	A failed adit led to numerous problems.
	Locate the efferent neuron.	Locate the afferent neuron.
	Globalization brought effluence to China	Globalization brought affluence to China
	Meet me at Ellen Street	Meet me at Allen Street
	It can be challenging to induce energy.	It can be challenging to induce anergy.
	Teenagers may encounter a social enemy.	Teenagers may encounter a social anomy.
	Meet me at Epple Road	Meet me at Apple Road
	Look at the genes and exons	Look at the genes and axons
	Look for chromosome bending.	Look for chromosome banding
	Calculate the betting averages	Calculate the batting averages
	Meet me at Kemble Road	Meet me at Campbell Road
	It will be an expensive study	It will be an expansive study
	Meet me at Fenn Street	Meet me at Fann Street
	She’s a little fleshy	She’s a little flashy
	I’d like to find a shop that sells gems	I’d like to find a shop that sells jams
	Meet me at Grenville Place	Meet me at Granville Place
	The athlete leapt everyone	The athlete lapped everyone
	Take a biopsy of that mess	Take a biopsy of that mass
	I’d like a pedal board for my birthday	I’d like a paddle board for my birthday
	Critics penned several recent articles	Critics panned several recent articles
	I just wrecked the pool table	I just racked the pool table
	Meet me at Redcliffe Square	Meet me at Radcliffe Square
	I don’t think renting is a good idea	I don’t think ranting is a good idea
	Good coffee requires a good temper	Good coffee requires a good tamper

Mono- and disyllabic word lists with similar consonantal neighbours

List of monosyllabic and disyllabic words with similar consonantal neighbours

Monosyllabic	Disyllabic
Band	Banding
bat	Batty
bend	Bending
bet	Betty
feel	Feeling
fill	Filling
leave	Leaving
live	Living
pack	Packing
pat	Patter
peck	Pecking
pet	Petter
sand	Sander
send	Sender
tamp	Tamper
temp	Temper
wheel	Wheelton
will	Whilton

Di- and trisyllabic word lists with similar consonantal neighbours

List of disyllabic and trisyllabic words with similar consonantal neighbours

Disyllabic	Trisyllabic
Expand	Expansive
Expend	Expensive

Individual vowel and syllable matrices

Creating a performance matrix with vowels and syllables illustrated a striking pattern. The odd word was more readily identified in single syllable words, and less well identified in multisyllabic words.

Vowels which had no direct L1 category to assimilate to (/ɪ/, /æ/) were easier for participants to identify as “odd” than vowels which had an equivalent L1 category (/i/, /ɛ/).

1.

Band (1.0)	Banding (.7)
Bend (.7)	Bending (.6)

2.

Bat (1.0)	Batty (1.0)
Bet (.7)	Betty (.8)

3.

Pack (1.0)	Packing (.9)
Peck (.8)	Pecking (.7)

4.

Sand (.9)	Sander (.3)
Send (.7)	Sender (.5)

5.

Will (1.0)	Whilton (.7)
Wheel (.9)	Wheelton (.2)

6.

Live (1.0)	Liver (.8)
Leave (.9)	Lever (.9)

7.

Fill (.5)	Filling (.7)
Feel (.7)	Feeling (.6)

8.

Tamp (1.0)	Tamper (.2)
Temp (1.0)	Temper (.3)

9.

Expand (.4)	Expansive (.5)
Expend (1.0)	Expensive (.2)

Pilot item analysis report

Item Analysis: /i , ɪ/ bVd

	Participant Raw Scores													α if
Item	01	02	03	04	05	06	07	08	09	10	Total	*P*	*rpbis*	Deleted
bead_key2_n	0	1	1	1	1	0	1	1	0	1	7	0.7	0.91	0.92
bid_key1_n	0	1	1	1	1	0	1	1	0	1	7	0.7	0.91	0.92

bead_key3_an	0	1	1	1	1	0	1	0	0	1	6	0.6	0.85	0.93
bead_key3_n	0	1	1	1	1	0	1	0	0	1	6	0.6	0.85	0.93
bid_key3_al	0	1	1	1	0	0	1	1	0	1	6	0.6	0.82	0.93
bead_key4_an	1	1	1	1	1	0	1	1	0	1	8	0.8	0.75	0.93
bid_key2_n	1	1	1	1	1	0	1	1	0	1	8	0.8	0.75	0.93
bead_key2_an	0	0	1	1	1	0	1	1	0	1	6	0.6	0.73	0.93
bead_key3_al	0	0	1	1	1	0	1	1	0	1	6	0.6	0.73	0.93
bead_key4_al	0	1	1	1	1	1	1	1	0	1	8	0.8	0.71	0.93
bid_key1_an	0	1	1	1	1	1	1	1	0	1	8	0.8	0.71	0.93
bead_key2_al	0	1	0	1	1	0	1	0	0	1	5	0.5	0.70	0.93
bid_key4_an	0	1	0	1	0	0	1	0	0	1	4	0.4	0.69	0.93
bead_key4_n	0	1	1	1	1	1	1	0	0	1	7	0.7	0.62	0.93
bead_key1_an	0	1	1	1	0	1	1	1	0	1	7	0.7	0.59	0.93
bead_key1_n	1	1	1	1	0	0	1	0	0	1	6	0.6	0.58	0.93
bid_key2_al	1	1	1	1	1	1	1	1	0	1	9	0.9	0.58	0.93
bid_key3_an	1	1	1	1	1	1	1	1	0	1	9	0.9	0.58	0.93
bid_key2_an	0	0	1	1	0	1	1	1	0	1	6	0.6	0.43	0.93
bead_key3_g	1	1	1	1	1	0	1	1	1	1	9	0.9	0.39	0.93
bid_key1_al	1	1	1	0	0	0	1	1	0	1	6	0.6	0.37	0.93
bead_key4_g	0	1	1	1	1	1	1	1	1	1	9	0.9	0.34	0.93
bid_key2_g	0	1	1	1	1	1	1	1	1	1	9	0.9	0.34	0.93
bid_key4_n	1	1	1	1	0	1	1	0	0	1	7	0.7	0.34	0.93
bead_key1_al	1	0	0	1	0	0	1	1	0	1	5	0.5	0.32	0.93
bid_key4_al	1	1	1	1	1	0	1	0	1	1	8	0.8	0.28	0.93
bid_key1_g	1	1	0	1	1	0	1	1	1	1	8	0.8	0.11	0.93
bead_key1_g	0	1	1	0	1	1	1	1	1	1	8	0.8	0.01	0.94
bid_key3_n	1	0	1	1	1	0	0	0	1	1	6	0.6	- 0.10	0.94
bead_key2_g	1	1	1	1	1	1	1	1	1	1	10	1		
bid_key3_g	1	1	1	1	1	1	1	1	1	1	10	1		
bid_key4_g	1	1	1	1	1	1	1	1	1	1	10	1		
Participant Total	**15**	**27**	**28**	**30**	**24**	**14**	**31**	**23**	**10**	**32**	**234**			
Participant Mean	**0.47**	**0.84**	**0.88**	**0.94**	**0.75**	**0.44**	**0.97**	**0.72**	**0.31**	**1.00**	**7.31**	**0.73**	**0.55**	

Item Analysis: /i , ɪ/ Diverse Words

Item	Participant Raw Scores													α if
	01	02	03	04	05	06	07	08	09	10	Total	*P*	*rpbis*	Deleted
beads_key3_n	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
carotene_key4_n	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
cheaper_key4_g	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
clique_key3_an	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
dip_key4_an	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
fickle_key4_al	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
leap_key2_n	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
leave_key3_g	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
lever_key3_al	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
licking_key1_g	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
litre_key3_n	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
peak_key3_g	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
pill_key2_n	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
pitter_key3_al	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
wheat_key4_al	1	1	1	1	1	0	1	1	1	1	9	0.9	0.88	0.95
deep_key2_an	1	1	1	1	1	0	1	1	1	0	8	0.8	0.78	0.95
lip_key3_n	1	1	1	1	1	0	1	1	1	0	8	0.8	0.78	0.95
peach_key4_al	1	1	1	1	1	0	1	1	1	0	8	0.8	0.78	0.95
scream_key3_al	1	1	1	1	1	0	1	1	1	0	8	0.8	0.78	0.95
seak_key4_g_b	1	1	1	1	1	0	1	1	1	0	8	0.8	0.78	0.95
whiz_key2_al	1	1	1	1	1	0	1	1	1	0	8	0.8	0.78	0.95
rich_key1_an	1	1	1	1	0	0	1	1	1	0	7	0.7	0.74	0.95
deacons_key2_n	0	1	1	1	0	0	1	1	1	0	6	0.6	0.72	0.95
seal_key1_n	0	1	1	1	0	0	1	1	1	0	6	0.6	0.72	0.95
deem_key4_al	1	1	1	1	0	0	1	1	1	1	8	0.8	0.72	0.95
keeper_key1_n	1	1	1	1	0	0	1	1	1	1	8	0.8	0.72	0.95
kipper_key3_al	0	1	1	1	1	0	1	1	1	0	7	0.7	0.71	0.95
litter_key2_g	0	1	1	1	1	0	1	1	1	1	8	0.8	0.69	0.95
liver_key2_al	0	1	1	1	1	0	1	1	1	1	8	0.8	0.69	0.95
wheeze_key3_an	0	1	1	1	1	0	1	1	1	1	8	0.8	0.69	0.95
beat_key2_al	1	0	1	1	0	0	1	1	1	0	6	0.6	0.66	0.95
fizz_key3_n	1	0	1	1	1	0	1	1	1	0	7	0.7	0.64	0.95
sleet_key2_al	1	0	1	1	1	0	1	1	1	0	7	0.7	0.64	0.95
leaking_key2_g	0	0	1	1	0	0	1	1	0	0	4	0.4	0.62	0.95
reason_key2_an	0	0	1	1	0	0	1	1	0	0	4	0.4	0.62	0.95
scrim_key3_an	1	0	1	1	1	0	1	1	1	1	8	0.8	0.61	0.95
siemens_road_key4_n	0	1	1	1	1	0	1	1	0	0	6	0.6	0.59	0.95
skim_key4_an	1	0	1	1	0	0	1	1	1	1	7	0.7	0.59	0.95
witt_key1_n	0	0	1	1	0	0	1	1	1	1	6	0.6	0.58	0.95
fist_key2_g	1	1	1	1	1	0	1	1	0	1	8	0.8	0.57	0.95
hit_key3_n	1	1	1	1	1	0	1	1	0	1	8	0.8	0.57	0.95

lever_key3_al_b	1	1	1	1	1	0	1	1	0	1	8	0.8	0.57	0.95
pitch_key2_an	1	1	1	1	1	0	1	1	0	1	8	0.8	0.57	0.95
win_key2_n	1	1	1	1	1	0	1	1	0	1	8	0.8	0.57	0.95
grid_key2_al	0	1	1	1	0	0	1	1	0	1	6	0.6	0.54	0.95
teeny_key1_g	0	1	1	1	0	0	1	1	0	1	6	0.6	0.54	0.95
sheep_lane_key2_an	0	1	0	1	0	0	1	1	0	0	4	0.4	0.52	0.95
filling_key3_an	0	1	1	1	1	0	1	1	0	1	7	0.7	0.52	0.95
eel_key3_an	0	1	1	1	0	0	1	0	1	0	5	0.5	0.50	0.95
sleeping_key2_al	1	1	0	1	1	0	1	1	1	1	8	0.8	0.49	0.95
slit_key1_n	1	1	0	1	1	0	1	1	1	1	8	0.8	0.49	0.95
heel_key2_g	1	0	0	1	1	0	1	1	1	0	6	0.6	0.47	0.95
slick_key1_al	1	1	1	0	1	0	1	1	1	1	8	0.8	0.46	0.95
sleek_key1_g	1	1	1	1	1	0	1	0	1	0	7	0.7	0.46	0.95
rip_key4_an	1	0	1	0	1	0	1	1	1	0	6	0.6	0.45	0.95
feel_key1_an	0	1	1	0	1	0	1	1	1	1	7	0.7	0.43	0.95
whitfield_key3_al	0	1	1	0	1	0	1	1	1	1	7	0.7	0.43	0.95
hip_key1_al	1	1	1	1	1	0	0	1	1	1	8	0.8	0.43	0.95
hymn_key4_an	1	1	1	1	1	0	0	1	1	1	8	0.8	0.43	0.95
feast_key2_g	1	0	1	1	1	0	0	1	1	0	6	0.6	0.42	0.95
greed_key3_g	0	1	1	1	1	0	0	1	1	1	7	0.7	0.40	0.95
scheme_key1_an	1	1	0	1	0	0	0	1	1	0	5	0.5	0.37	0.95
fecal_key1_al	0	1	1	1	0	0	0	1	0	1	5	0.5	0.36	0.95
tinny_key1_al	1	1	0	1	1	0	1	1	0	1	7	0.7	0.35	0.95
simmons_road_key3_n	0	0	1	1	0	0	1	0	0	1	4	0.4	0.31	0.95
wheelton_key2_al	0	0	0	0	0	0	1	0	1	0	2	0.2	0.27	0.95
heme_key2_an	1	0	1	0	0	0	1	0	0	0	3	0.3	0.26	0.95
reach_key3_an	1	1	0	1	1	0	1	0	1	1	7	0.7	0.21	0.95
peter_key1_al	0	1	1	1	0	1	1	1	1	0	7	0.7	0.18	0.95
been_key3_g	1	1	1	1	1	1	1	1	1	0	9	0.9	0.15	0.95
pick_key1_g	1	1	1	1	1	1	1	1	1	0	9	0.9	0.15	0.95
weak_key2_an	1	0	1	0	0	0	0	0	1	0	3	0.3	0.14	0.95
fill_key2_an	1	1	0	0	0	0	1	0	1	1	5	0.5	0.10	0.96
scene_key1_n	1	1	1	1	0	1	1	1	0	0	7	0.7	0.08	0.96
sill_key4_n	1	1	1	1	0	1	1	1	1	1	9	0.9	0.08	0.95
bids_key2_g	0	0	1	1	0	1	1	1	1	1	7	0.7	0.04	0.96
dickens_key1_n	0	1	1	1	1	1	1	1	1	1	9	0.9	0.04	0.95
ill_key1_al	0	1	1	1	1	1	1	1	1	1	9	0.9	0.04	0.95
seep_key3_n	0	1	1	1	1	1	1	1	1	1	9	0.9	0.04	0.95
ween_key1_g	0	1	1	1	1	1	1	1	1	1	9	0.9	0.04	0.95
risen_key1_n	0	0	0	0	0	0	0	0	0	0	0	0	0.00	
click_key3_an	1	1	0	1	1	1	1	1	1	0	8	0.8	-0.05	0.96
gin_key1_g	1	0	1	1	1	1	1	1	1	1	9	0.9	-0.06	0.96
sick_key3_g	1	0	1	1	1	1	1	1	1	1	9	0.9	-0.06	0.96

mill_key3_al	1	1	1	1	1	1	1	1	0	1	9	0.9	-0.12	0.96
feeling_key1_an	0	1	0	1	0	1	1	1	0	1	6	0.6	-0.13	0.96
cheak_key1_al	1	0	0	1	0	1	1	1	1	1	7	0.7	-0.13	0.96
whilton_key1_an	1	0	0	1	0	1	1	1	1	1	7	0.7	-0.13	0.96
ship_lane_key4_an	1	1	0	0	0	1	1	1	1	0	6	0.6	-0.14	0.96
sin_key2_g	1	1	1	1	0	1	0	1	1	1	8	0.8	-0.17	0.96
wheel_key4_g	1	1	1	0	1	1	1	1	1	1	9	0.9	-0.26	0.96
heap_key3_al	1	0	1	0	1	1	0	1	1	0	6	0.6	-0.27	0.96
fees_key2_n	1	1	1	1	1	1	0	0	1	0	7	0.7	-0.31	0.96
chick_key1_an	1	0	0	0	1	1	1	1	0	1	6	0.6	-0.39	0.96
reap_key4_g	1	0	1	0	1	1	0	1	1	1	7	0.7	-0.40	0.96
bin_key1_g	0	0	0	0	0	0	0	0	0	0	0	0		
bit_key4_al	1	1	1	1	1	1	1	1	1	1	10	1		
chipper_key3_g	1	1	1	1	1	1	1	1	1	1	10	1		
dim_key4_al	1	1	1	1	1	1	1	1	1	1	10	1		
gene_key1_g	1	1	1	1	1	1	1	1	1	1	10	1		
heat_key4_n	1	1	1	1	1	1	1	1	1	1	10	1		
hid_key1_g	1	1	1	1	1	1	1	1	1	1	10	1		
hill_key4_g	1	1	1	1	1	1	1	1	1	1	10	1		
lead_key4_an	1	1	1	1	1	1	1	1	1	1	10	1		
lid_key2_n	1	1	1	1	1	1	1	1	1	1	10	1		
live_key3_an	1	1	1	1	1	1	1	1	1	1	10	1		
meal_key4_an	1	1	1	1	1	1	1	1	1	1	10	1		
peel_key4_al	1	1	1	1	1	1	1	1	1	1	10	1		
seak_key4_g	1	1	1	1	1	1	1	1	1	1	10	1		
sip_key4_al	1	1	1	1	1	1	1	1	1	1	10	1		
slipping_key4_n	1	1	1	1	1	1	1	1	1	1	10	1		
wheatfield_key4_n	1	1	1	1	1	1	1	1	1	1	10	1		
wick_key4_n	1	1	1	1	1	1	1	1	1	1	10	1		
will_key3_g	1	1	1	1	1	1	1	1	1	1	10	1		
Participant Total	**84**	**89**	**97**	**99**	**82**	**42**	**101**	**103**	**92**	**78**	**867**			
Participant Mean	**0.74**	**0.78**	**0.85**	**0.87**	**0.72**	**0.37**	**0.89**	**0.90**	**0.81**	**0.68**	**7.61**	**0.76**	**0.43**	

Item Analysis: /i , ɪ/ Sentences

	Participant Raw Scores												
Item	01	02	03	04	05	06	07	08	09	10	Total	*P*	*rpbis*	α if Deleted
s_bids_key1_g	0	1	0	1	0	0	1	1	1	0	5	0.5	0.90	0.80
s_tinny_key1_n	0	1	0	0	0	0	1	1	0	0	3	0.3	0.73	0.81
s_beaten_key3_al	0	1	1	1	0	0	1	1	1	0	6	0.6	0.70	0.81
s_slit_key3_an	0	1	1	1	0	0	1	1	1	0	6	0.6	0.70	0.81
s_lid_key1_an	1	1	0	1	1	0	1	1	1	0	7	0.7	0.68	0.81
s_peel_key4_n	1	1	0	1	1	0	1	1	1	0	7	0.7	0.68	0.81
s_wheatfield_key4_al	1	1	0	1	0	0	1	1	1	1	7	0.7	0.68	0.81
s_teeny_key1_n	0	1	0	0	0	0	1	0	0	0	2	0.2	0.66	0.81
s_sill_key2_an	0	1	0	1	0	0	0	1	0	0	3	0.3	0.64	0.81
s_slipping_key1_al	0	0	0	1	0	0	1	1	0	0	3	0.3	0.64	0.81
s_cheaper_key3_n	0	1	0	1	0	0	0	1	1	0	4	0.4	0.61	0.81
s_dickens_key2_n	0	1	0	1	0	0	0	1	1	0	4	0.4	0.61	0.81
s_sheep_lane_key1_n	1	1	0	0	0	0	1	1	1	0	5	0.5	0.58	0.81
s_sleet_key2_al	0	1	1	1	1	0	1	1	1	0	7	0.7	0.56	0.81
s_feast_key1_al	0	0	0	1	0	0	1	0	0	0	2	0.2	0.56	0.82
s_heat_key4_an	0	0	0	1	0	0	1	0	0	0	2	0.2	0.56	0.82
s_deacons_key4_al	0	1	0	1	0	0	1	0	1	1	5	0.5	0.55	0.81
s_lead_key1_g	1	1	0	1	0	1	1	1	1	0	7	0.7	0.53	0.81
s_scenes_key1_an	0	1	0	0	0	0	1	1	1	1	5	0.5	0.50	0.82
s_beads_key2_an	0	1	0	0	0	0	1	1	0	1	4	0.4	0.49	0.82
s_fickle_key1_al	0	0	0	0	0	0	1	1	0	0	2	0.2	0.49	0.82
s_bitten_key2_g	1	1	1	1	1	0	1	1	1	1	9	0.9	0.46	0.82
s_bidding_key4_al	0	1	1	1	1	0	1	1	1	1	8	0.8	0.44	0.82
s_ship_lane_key3_al	0	1	1	1	1	0	1	1	1	1	8	0.8	0.44	0.82
s_feel_key3_al	0	1	0	0	0	0	0	0	0	0	1	0.1	0.43	0.82
s_leave_key4_n	1	1	0	1	0	1	1	1	1	1	8	0.8	0.41	0.82
s_gins_key4_an	1	1	0	1	0	1	1	0	1	0	6	0.6	0.33	0.82
s_pill_key3_al	0	1	0	0	1	0	1	1	1	1	6	0.6	0.33	0.82
s_mills_key2_an	1	1	0	0	0	0	0	1	0	0	3	0.3	0.31	0.82
s_live_key2_an	0	1	1	0	0	0	1	1	1	1	6	0.6	0.30	0.82
s_simmons_road_key1_n	0	0	0	1	0	0	0	0	0	0	1	0.1	0.29	0.82
s_kipper_key4_al	1	1	0	1	1	1	1	1	1	1	9	0.9	0.29	0.82
s_hit_key1_g	0	1	0	1	0	0	0	0	0	1	3	0.3	0.28	0.82
s_fist_key3_g	1	1	1	1	1	1	1	1	1	0	9	0.9	0.24	0.82
s_wheezed_key3_g	0	1	0	0	1	1	1	1	0	0	5	0.5	0.17	0.82
s_carotene_key1_an	1	0	0	1	0	0	1	0	0	1	4	0.4	0.16	0.82
s_fecal_key2_g	1	1	0	1	1	0	0	0	0	0	4	0.4	0.16	0.82
s_scheming_key2_n	0	0	0	1	0	1	1	0	0	0	3	0.3	0.14	0.82
s_whizzed_key3_an	1	1	0	0	0	1	1	1	1	1	7	0.7	0.12	0.82
s_beading_key2_an	0	0	1	1	0	0	1	0	0	1	4	0.4	0.05	0.83
s_seal_key4_g	1	1	1	0	1	0	1	0	0	0	5	0.5	0.04	0.83

s_chipper_key1_an	0	0	1	1	1	0	1	1	0	1	6	0.6	0.03	0.83
s_keeper_key3_n	1	0	1	1	1	0	1	1	0	1	7	0.7	-0.05	0.83
s_wheelton_key2_n	1	1	1	0	1	0	0	1	1	0	6	0.6	-0.07	0.83
s_skimming_key2_al	1	1	0	1	1	1	0	0	0	0	5	0.5	-0.14	0.83
s_meals_key2_al	1	0	0	0	1	0	0	1	0	0	3	0.3	-0.18	0.83
s_sins_key2_g	1	1	0	0	0	1	0	0	0	0	3	0.3	-0.18	0.83
s_keratin_key3_g	1	1	1	1	1	0	0	0	0	1	6	0.6	-0.21	0.83
s_siemens_road_key4_an	1	0	1	1	1	1	1	0	1	0	7	0.7	-0.33	0.83
s_sleeping_key4_g	1	0	1	1	0	1	1	0	0	1	6	0.6	-0.33	0.84
s_fill_key3_n	0	0	1	1	0	1	0	1	1	1	6	0.6	-0.36	0.84
s_whilton_key4_n	0	0	0	0	0	1	0	0	1	0	2	0.2	-0.40	0.83
s_genes_key1_g	0	1	1	0	1	1	0	0	1	0	5	0.5	-0.42	0.84
s_pick_key3_n	0	0	1	0	0	0	0	0	0	1	2	0.2	-0.49	0.84
s_peek_key2_g	1	0	1	0	1	1	1	0	0	0	5	0.5	-0.52	0.84
s_whitfield_key4_an	1	1	1	1	1	1	1	1	1	1	10	1		
Participant Total	**25**	**39**	**21**	**36**	**22**	**17**	**39**	**34**	**29**	**22**	**284**			
Participant Mean	**0.45**	**0.70**	**0.38**	**0.64**	**0.39**	**0.30**	**0.70**	**0.61**	**0.52**	**0.39**	**5.07**	**0.51**	**0.27**	

Item Analysis: /ɛ, æ/ bVd

	Participant Raw Scores													α if
Item	01	02	03	04	05	06	07	08	09	10	Total	*P*	*rpbis*	Deleted
bed_key2_g	1	0	0	1	1	0	1	0	1	1	6	0.6	0.77	0.88
bad_key2_al	1	1	0	1	0	0	1	0	1	1	6	0.6	0.74	0.88
bed_key1_g	1	1	0	1	0	0	1	0	1	1	6	0.6	0.74	0.88
bad_key3_an	1	0	0	1	1	0	0	0	1	1	5	0.5	0.71	0.88
bed_key1_an	1	1	0	1	1	0	1	1	1	1	8	0.8	0.65	0.89
bed_key2_n	0	0	0	1	0	0	1	1	1	1	5	0.5	0.64	0.89
bad_key3_al	1	0	0	1	1	0	0	1	1	1	6	0.6	0.63	0.89
bad_key4_an	1	0	0	1	1	0	0	1	1	1	6	0.6	0.63	0.89
bad_key2_g	0	1	0	1	0	0	1	0	1	1	5	0.5	0.57	0.89
bed_key4_g	1	0	1	1	1	0	1	0	1	1	7	0.7	0.54	0.89
bad_key1_g	1	0	0	0	0	0	1	1	1	1	5	0.5	0.50	0.89
bed_key1_al	1	1	1	1	0	0	1	1	1	1	8	0.8	0.48	0.89
bed_key4_n	1	1	1	1	0	0	1	1	1	1	8	0.8	0.48	0.89
bed_key1_n	1	1	1	1	1	0	1	1	1	1	9	0.9	0.47	0.89
bed_key3_g	1	0	1	1	1	0	0	0	1	1	6	0.6	0.46	0.89
bad_key3_g	1	0	0	1	1	1	1	1	1	1	8	0.8	0.44	0.89
bad_key4_g	1	0	0	1	1	1	1	1	1	1	8	0.8	0.44	0.89
bed_key2_an	1	1	1	1	1	0	1	0	1	1	8	0.8	0.44	0.89
bed_key4_al	1	1	0	1	0	1	1	0	1	1	7	0.7	0.43	0.89
bad_key1_al	1	1	0	1	0	0	0	1	1	0	5	0.5	0.37	0.89
bed_key2_al	0	1	0	1	0	0	0	0	1	0	3	0.3	0.37	0.89
bad_key2_n	1	1	0	1	1	1	1	1	1	1	9	0.9	0.37	0.89
bad_key2_an	1	1	0	1	1	0	0	0	1	0	5	0.5	0.34	0.89
bed_key3_n	1	0	1	1	0	1	1	1	1	1	8	0.8	0.28	0.89
bad_key3_n	1	0	1	1	1	1	1	1	1	1	9	0.9	0.21	0.89
bed_key3_an	1	0	1	1	1	1	1	1	1	1	9	0.9	0.21	0.89
bad_key1_an	1	0	1	1	0	1	0	1	1	1	7	0.7	0.18	0.90
bad_key4_al	1	1	1	1	0	1	1	1	1	1	9	0.9	0.15	0.90
bed_key3_al	0	1	1	1	1	1	1	0	1	1	8	0.8	- 0.12	0.90
bad_key4_n	0	1	1	1	1	1	1	1	1	1	9	0.9	- 0.26	0.90
bad_key1_n	1	1	1	1	1	1	1	1	1	1	10	1		
bed_key4_an	1	1	1	1	1	1	1	1	1	1	10	1		
Participant Total	**27**	**18**	**15**	**31**	**19**	**13**	**24**	**20**	**32**	**29**	**228**			
Participant Mean	**0.84**	**0.56**	**0.47**	**0.97**	**0.59**	**0.41**	**0.75**	**0.63**	**1.00**	**0.91**	**7.13**	**0.71**	**0.43**	

Item Analysis: /ɛ, æ/ Diverse Words

	Participant Raw Scores													α if
Item	01	02	03	04	05	06	07	08	09	10	Total	*P*	*rpbis*	Deleted
epple_road_key2_an	1	1	0	1	0	0	1	1	1	0	6	0.6	0.81	0.93
flash_key2_an	1	1	0	1	1	0	1	1	1	0	7	0.7	0.78	0.93
track_key2_n	1	1	0	1	1	0	1	1	1	1	8	0.8	0.71	0.93
expand_key4_al	0	1	0	1	0	0	0	1	1	0	4	0.4	0.69	0.93
adapt_key3_an	0	0	0	1	0	0	0	0	1	0	2	0.2	0.68	0.93
panned_key1_an	1	0	0	1	0	0	0	0	1	0	3	0.3	0.67	0.93
blend_key2_g	1	1	1	1	0	0	1	1	1	0	7	0.7	0.67	0.93
gas_key3_n	1	1	1	1	1	0	1	1	1	0	8	0.8	0.67	0.93
kettle_key1_b	1	1	1	1	1	0	1	1	1	0	8	0.8	0.67	0.93
pan_key3_n	1	1	1	1	1	0	1	1	1	0	8	0.8	0.67	0.93
acton_key2_al	1	1	1	1	1	0	1	1	1	1	9	0.9	0.66	0.93
bending_key3_an	1	1	1	1	1	0	1	1	1	1	9	0.9	0.66	0.93
fetter_key1_an	1	1	1	1	1	0	1	1	1	1	9	0.9	0.66	0.93
gem_key4_al	1	1	1	1	1	0	1	1	1	1	9	0.9	0.66	0.93
paddle_key3_g	1	1	1	1	1	0	1	1	1	1	9	0.9	0.66	0.93
guess_key1_al	1	0	0	1	0	0	0	1	1	0	4	0.4	0.64	0.93
redcliffe_square_key2_al	1	0	0	1	0	0	0	1	1	0	4	0.4	0.64	0.93
aster_key2_an	0	1	0	1	1	0	1	1	1	0	6	0.6	0.64	0.93
leapt_key1_al	1	1	0	1	1	0	1	0	1	1	7	0.7	0.59	0.93
mend_key4_g	1	1	0	1	0	0	0	0	1	1	5	0.5	0.59	0.93
campbell_road_key2_an	0	0	0	1	0	0	1	1	1	0	4	0.4	0.59	0.93
lend_key3_g	0	1	0	1	1	0	0	0	1	0	4	0.4	0.59	0.93
betty_key1_an	1	1	1	1	0	0	1	1	1	1	8	0.8	0.58	0.93
ham_key2_g	1	1	1	1	0	0	1	1	1	1	8	0.8	0.58	0.93
rant_key1_n	1	1	1	1	0	0	1	1	1	1	8	0.8	0.58	0.93
bet_key2_al	1	1	0	1	1	0	0	1	1	1	7	0.7	0.57	0.93
patter_key1_an	1	1	0	1	1	0	0	1	1	1	7	0.7	0.57	0.93
men_key2_al	1	0	0	1	1	0	1	0	1	0	5	0.5	0.56	0.93
beck_key3_al	1	1	1	1	1	0	1	0	1	0	7	0.7	0.56	0.93
ecton_key2_an	1	1	1	1	1	0	1	0	1	0	7	0.7	0.56	0.93
effluent_key1_al	1	0	0	1	1	0	0	0	1	0	4	0.4	0.54	0.93
land_key1_an	1	0	0	1	0	0	1	1	1	1	6	0.6	0.52	0.93
affluent_key3_an	1	1	0	1	0	0	0	1	0	0	4	0.4	0.50	0.93
pedal_key2_an	1	0	0	1	1	0	1	1	1	1	7	0.7	0.47	0.93
kemble_road_key2_al	1	0	0	1	0	0	0	0	0	0	2	0.2	0.45	0.93
cattle_key1_n	1	1	0	1	1	0	1	1	0	0	6	0.6	0.45	0.93
peck_key1_an	1	1	0	1	1	1	1	1	1	0	8	0.8	0.38	0.93
ester_key2_n	1	0	0	0	0	0	1	0	1	0	3	0.3	0.37	0.93
ellen_key2_g	1	1	0	0	1	0	0	1	1	0	5	0.5	0.36	0.93
fatter_key1_g	1	1	1	1	0	0	1	1	0	0	6	0.6	0.35	0.93
granville_place_key3_al	0	1	1	1	0	0	0	0	1	1	5	0.5	0.32	0.93

trek_key3_n	1	0	1	1	1	0	1	1	1	1	8	0.8	0.32	0.93
bend_key3_g	0	1	1	1	1	0	1	0	1	1	7	0.7	0.30	0.93
addison_road_key2_n	0	1	0	1	0	1	1	1	1	0	6	0.6	0.30	0.93
effluence_key3_n	0	1	0	1	0	0	1	1	0	1	5	0.5	0.29	0.93
edison_road_key3_n	0	1	0	1	0	0	1	0	0	1	4	0.4	0.28	0.93
hem_key4_n	0	0	1	1	1	0	0	1	1	0	5	0.5	0.28	0.93
pack_key2_g	1	1	0	1	1	1	1	1	1	1	9	0.9	0.27	0.93
grenville_place_key4_alison	0	1	1	1	1	0	0	0	1	1	6	0.6	0.25	0.93
rent_key3_an	0	1	0	1	1	1	1	1	1	0	7	0.7	0.23	0.93
pet_key2_n	1	0	0	1	0	1	1	0	1	0	5	0.5	0.23	0.93
essay_key1_g	0	1	0	0	0	0	0	0	1	1	3	0.3	0.22	0.93
packing_key4_al	1	1	1	1	1	1	1	1	1	0	9	0.9	0.22	0.93
pallet_key2_al	1	1	0	1	0	1	0	1	1	1	7	0.7	0.22	0.93
apple_road_key2_al	0	1	1	0	0	0	1	1	1	0	5	0.5	0.21	0.93
fann_street_key1_n	1	0	0	1	1	1	0	0	1	0	5	0.5	0.11	0.93
affluence_key1_al	1	1	1	1	0	1	1	1	1	1	9	0.9	0.11	0.93
mass_key1_al	0	1	0	1	1	1	1	1	1	1	8	0.8	0.09	0.93
hatley_road_key3_an	0	0	1	0	0	0	0	0	1	0	2	0.2	0.06	0.93
adept_key3_g	0	1	0	1	0	1	0	0	0	0	3	0.3	-0.01	0.93
allen_key2_g	0	1	1	1	0	1	1	1	1	1	8	0.8	-0.03	0.93
radcliffe_square_key2_an	1	0	0	0	1	0	0	0	0	0	2	0.2	-0.04	0.93
lapped_key4_al	1	1	1	1	1	1	1	0	1	1	9	0.9	-0.05	0.93
penned_key4_g	1	1	1	1	1	1	1	0	1	1	9	0.9	-0.05	0.93
petter_key4_al	1	1	1	1	1	1	1	0	1	1	9	0.9	-0.05	0.93
sand_key2_n	1	1	1	1	1	1	0	1	1	1	9	0.9	-0.08	0.93
pecking_key4_n	1	0	1	1	0	1	1	0	1	1	7	0.7	-0.10	0.93
wreck_key1_n	1	1	1	0	0	0	0	1	0	1	5	0.5	-0.14	0.93
pecked_key3_al	0	1	1	1	1	1	1	1	1	1	9	0.9	-0.15	0.93
rack_key2_g	0	1	1	1	1	1	1	1	1	1	9	0.9	-0.15	0.93
assay_key1_n	1	1	1	1	1	1	1	1	0	1	9	0.9	-0.44	0.93
axon_key4_al	1	1	1	1	1	1	1	1	1	1	10	1		
back_key4_n	1	1	1	1	1	1	1	1	1	1	10	1		
band_key4_g	1	1	1	1	1	1	1	1	1	1	10	1		
bat_key1_g	1	1	1	1	1	1	1	1	1	1	10	1		
batty_key4_n	1	1	1	1	1	1	1	1	1	1	10	1		
bland_key3_al	1	1	1	1	1	1	1	1	1	1	10	1		
expend_key4_an	1	1	1	1	1	1	1	1	1	1	10	1		

fenn_street_key4_g	1	1	1	1	1	1	1	1	1	1	10	1	
flesh_key4_an	1	1	1	1	1	1	1	1	1	1	10	1	
had_key4_n	1	1	1	1	1	1	1	1	1	1	10	1	
head_key3_an	1	1	1	1	1	1	1	1	1	1	10	1	
jam_key4_n	1	1	1	1	1	1	1	1	1	1	10	1	
kant_key3_g	1	1	1	1	1	1	1	1	1	1	10	1	
man_key3_g	1	1	1	1	1	1	1	1	1	1	10	1	
manned_key3_al	1	1	1	1	1	1	1	1	1	1	10	1	
mess_key1_an	1	1	1	1	1	1	1	1	1	1	10	1	
packed_key4_g	1	1	1	1	1	1	1	1	1	1	10	1	
pen_key4_an	1	1	1	1	1	1	1	1	1	1	10	1	
tamp_key1_g	1	1	1	1	1	1	1	1	1	1	10	1	
Participant Total	**69**	**72**	**53**	**83**	**59**	**39**	**66**	**65**	**80**	**55**	**641**		
Participant Mean	**0.77**	**0.80**	**0.59**	**0.92**	**0.66**	**0.43**	**0.73**	**0.72**	**0.89**	**0.61**	**7.12**	**0.71**	**0.37**

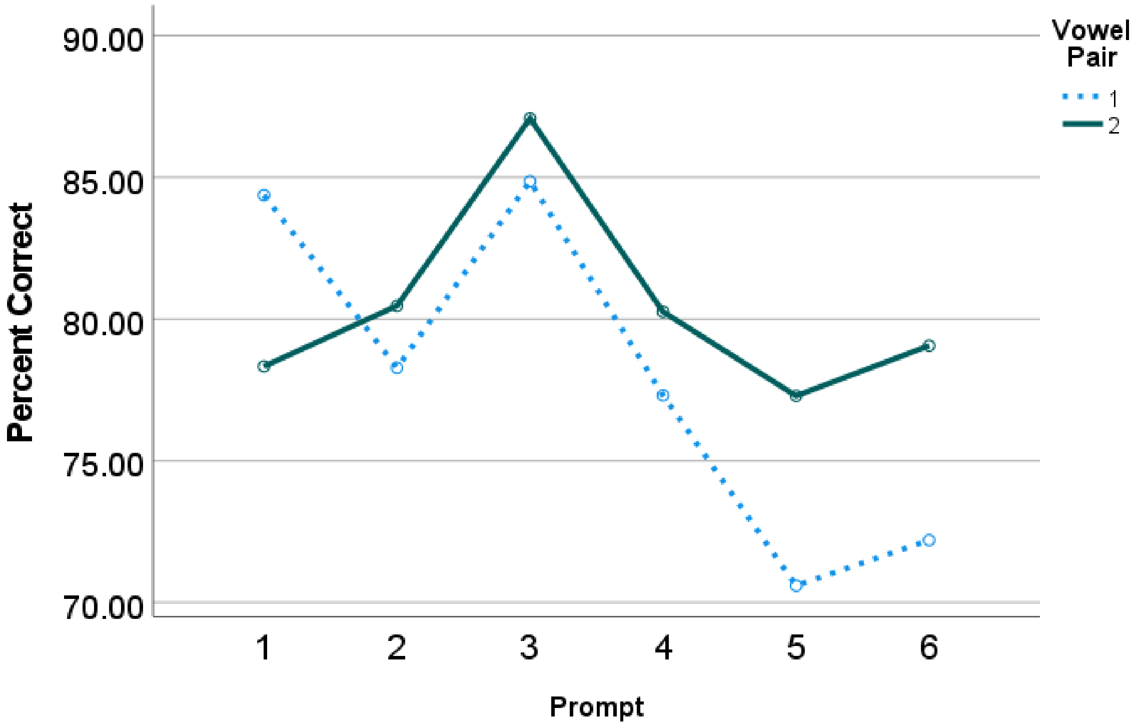
Item Analysis: /ɛ, æ/ Sentences

	Participant Raw Scores												α if
Item	01	02	03	04	05	06	07	08	09	10	Total	*P*	*rpbis*	Deleted
s_enemy_key1_an	0	1	0	1	0	0	0	0	1	0	3	0.3	0.85	0.78
s_renting_key2_al	0	0	0	1	0	0	0	0	1	0	2	0.2	0.82	0.78
s_acton_key3_g	0	0	0	1	0	0	0	0	0	0	1	0.1	0.63	0.79
s_anergy_key1_n	1	0	0	1	0	0	0	0	1	0	3	0.3	0.61	0.78
s_racked_key3_al	1	0	0	1	0	0	0	0	1	0	3	0.3	0.61	0.78
s_ecton_key4_an	0	1	0	1	0	1	0	1	1	0	5	0.5	0.58	0.78
s_fenn_street_key1_al	1	1	0	1	0	0	1	0	1	1	6	0.6	0.58	0.79
s_paddle_board_key4_al	0	1	0	1	1	0	1	1	1	0	6	0.6	0.58	0.79
s_axons_key1_an	0	0	0	1	1	0	0	0	1	0	3	0.3	0.57	0.79
s_panned_key1_n	0	1	0	1	0	1	0	0	1	1	5	0.5	0.55	0.79
s_gems_key4_an	1	1	0	1	1	1	1	1	1	1	9	0.9	0.53	0.79
s_granville_place_key4_g	1	1	0	1	1	1	1	1	1	1	9	0.9	0.53	0.79
s_bending_key1_n	1	0	0	1	0	0	0	1	1	0	4	0.4	0.51	0.79
s_ranting_key3_naina	0	1	0	1	0	1	1	0	1	1	6	0.6	0.51	0.79
s_mass_key4_g	1	1	0	1	0	1	1	1	1	1	8	0.8	0.50	0.79
s_expansive_key2_n	1	1	0	1	0	1	1	0	1	1	7	0.7	0.47	0.79
s_fleshy_key4_an	0	1	0	1	1	1	1	1	1	1	8	0.8	0.46	0.79
s_kemble_road_key3_al	0	1	0	0	0	0	0	0	1	0	2	0.2	0.46	0.79
s_batting_key1_g	0	0	0	1	0	1	1	0	1	0	4	0.4	0.45	0.79
s_radcliffe_square_key2_an	1	0	0	1	0	0	0	1	1	1	5	0.5	0.42	0.79
s_penned_key2_g	0	1	1	1	0	0	0	0	1	0	4	0.4	0.42	0.79
s_tamper_key1_g	1	0	0	1	1	0	0	1	1	0	5	0.5	0.39	0.79
s_addison_road_key2_n	0	1	0	1	0	0	1	1	0	0	4	0.4	0.39	0.79
s_betting_key4_al	0	0	0	1	1	1	0	0	1	0	4	0.4	0.39	0.79
s_afferent_key3_an	0	1	0	1	0	0	1	1	0	1	5	0.5	0.30	0.79
s_grenville_place_key1_al	1	1	0	1	0	0	0	1	0	1	5	0.5	0.27	0.80
s_adit_key2_al	0	0	0	1	0	0	1	0	0	1	3	0.3	0.24	0.80
s_energy_key2_n	0	0	0	0	0	0	1	0	1	0	2	0.2	0.24	0.80
s_lapped_key2_al	0	0	0	0	0	0	1	0	1	0	2	0.2	0.24	0.80
s_allen_street_key4_al	1	1	1	1	1	0	1	0	1	0	7	0.7	0.21	0.80
s_fann_street_key2_al	1	1	1	1	1	0	1	1	1	1	9	0.9	0.18	0.80
s_campbell_road_key3_g	0	1	1	1	0	1	1	0	1	1	7	0.7	0.17	0.80
s_peddle_board_key4_al	0	1	1	1	1	1	0	1	1	0	7	0.7	0.14	0.80
s_exons_key1_n	0	0	0	0	0	1	0	0	1	0	2	0.2	0.13	0.80
s_mess_key3_an	0	1	1	1	1	1	1	1	1	0	8	0.8	0.13	0.80
s_effluence_key4_n	1	1	0	1	1	1	0	1	0	0	6	0.6	0.11	0.80
s_flashy_key1_al	1	1	0	1	1	1	0	0	0	1	6	0.6	0.08	0.80
s_anomie_key4_g	0	0	1	1	1	0	1	0	1	1	6	0.6	0.05	0.80
s_wrecked_key1_an	1	0	0	0	1	0	0	0	1	0	3	0.3	0.05	0.80
s_redcliffe_square_key1_g	1	1	0	0	0	0	0	0	0	0	2	0.2	0.02	0.80
s_edison_road_key3_g	0	1	1	1	0	0	1	0	0	1	5	0.5	0.00	0.80
s_epple_road_key2_n	1	1	1	1	0	0	0	1	0	0	5	0.5	0.00	0.80
s_banding_key1_an	0	0	1	1	0	0	0	0	0	1	3	0.3	-0.08	0.81
s_efferent_key2_an	0	0	0	0	0	0	0	0	0	1	1	0.1	-0.16	0.80

s_edit_key2_n	0	0	0	0	1	0	0	0	0	0	1	0.1	-0.21	0.81
s_affluence_key4_n	1	0	0	0	1	1	0	1	1	1	6	0.6	-0.22	0.81
s_apple_road_key3_a	0	0	0	0	1	0	1	1	0	0	3	0.3	-0.27	0.81
s_expensive_key3_g_b	1	0	0	0	1	0	0	1	0	0	3	0.3	**-0.30**	**0.81**
s_expensive_key3_g	0	1	1	0	0	0	0	1	0	0	3	0.3	**-0.33**	**0.81**
s_jams_key2_an	1	0	0	0	1	1	0	1	0	0	4	0.4	-0.42	0.82
s_ellen_street_key3_g	0	1	1	0	0	1	1	0	0	1	5	0.5	-0.51	0.82
Participant Total	**21**	**28**	**12**	**37**	**20**	**19**	**22**	**22**	**33**	**21**	**235**			
Participant Mean	**0.41**	**0.55**	**0.24**	**0.73**	**0.39**	**0.37**	**0.43**	**0.43**	**0.65**	**0.41**	**4.61**	**0.46**	**0.25**	

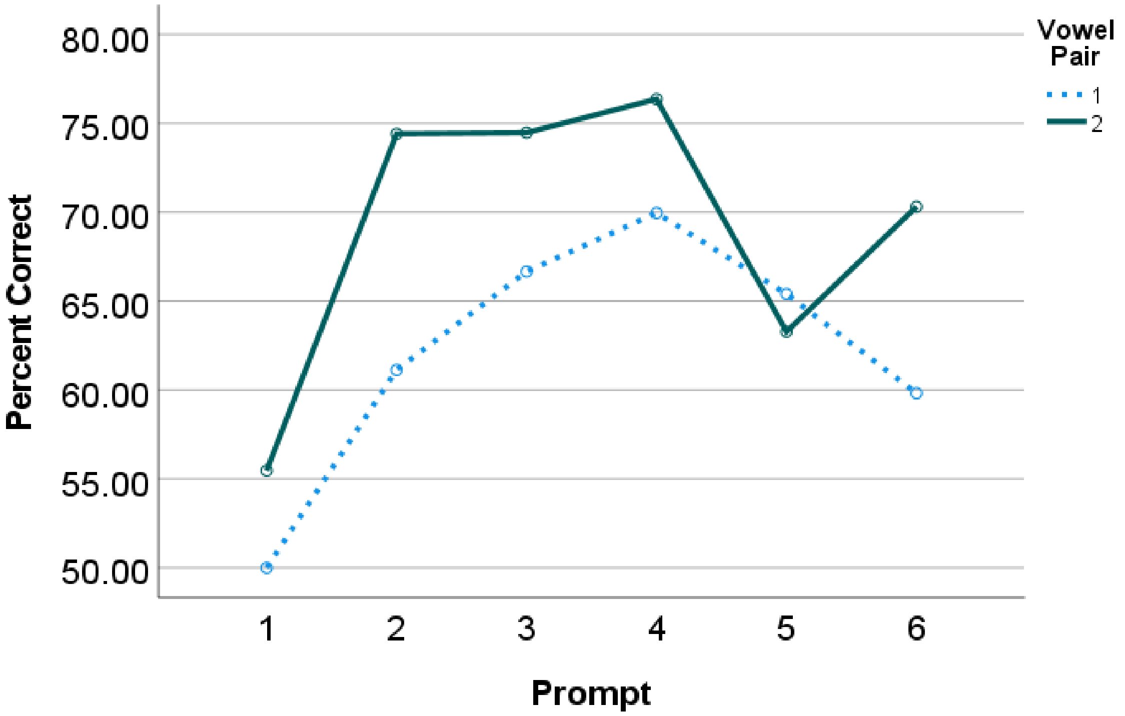
Experiment 1 Appendix

Mandarin two way (vowel pair and prompt) repeated measures ANOVA



Note. Vowel pair 1 = /ɛ, æ/; vowel pair 2 = /i, ɪ/. Prompt 1 = bVt Oddity; 2 = Diverse Words Oddity; 3 = bVt Identification; 4 = Diverse Words Identification; 5 = Directions; 6 = Diverse Sentences.

Korean two way (vowel pair and prompt) repeated measures ANOVA



Note. Vowel pair 1 = /ɛ, æ/; vowel pair 2 = /i, ɪ/. Prompt 1 = bVt Oddity; 2 = Diverse Words Oddity; 3 = bVt Identification; 4 = Diverse Words Identification; 5 = Directions; 6 = Diverse Sentences. A dependent *t*-test found the irregular finding in Directions was non-significant.

Isolated Words Vowel and Syllabicity Interaction

Results from the pilot suggested an interaction between vowel and syllable type for the oddity discrimination task. Experiment 1 did not show the same interaction in the oddity tasks, but did for the identification tasks in the high vowel pair.

For Experiment 1 oddity, a generalised linear mixed model analysis (outcome variable: correct response; fixed effects: vowel type, syllabicity; random effects: participant, item) revealed no significant interaction between syllabicity and vowel for either vowel pair. For /ɛ, æ/, confidence intervals showed the interaction was non-significant (*CI* = [.38, 1.48]), as was vowel type independently (*CI* = [.90, 2.06]); however, multisyllabicity was significant (*OR* = .61, *CI* = [.38, .98]), indicating multisyllabic words were more likely to yield an incorrect response compared with monosyllabic words. No significant finding was observed for the /i, ɪ/ vowel pair, whether for the interaction between vowel and syllabicity (*CI* = [.27, 4.48]), or for vowel (*CI* = [.32, 1.82]) or syllabicity independently (*CI* = [.21, 1.54]).

For Experiment 1 identification in the /i, ɪ/ vowel pair, the odds of obtaining a correct response were significantly greater when the target words was both multisyllabic and /ɪ/ (*OR* = 5.37, *CI* = [2.53, 11.41]). As with Sentences, this was interesting because individually, multisyllabic words (*OR* = .26 *CI* = [.15-.44]) and words with /ɪ/ (*OR* .45 *CI* = [.28, .72]) yielded lower scores.

In the mid-low vowel pair, the interaction between vowel and syllable type was non-significant (*CI* = [.38, 1.48]). Multisyllabicity was significant (*OR* = .61, *CI* = [.38, .98]) while vowel type was not (*CI* = [.90, 2.06]).

Mixed model outputs (excluding control)

Experiment 1 /i, ɪ/ Model comparison of prompt and language predictors for Diverse Sentences. Table shows model with all data (top) and an adjusted model with overlapping words removed (bottom).

					LRT		
Model	Fixed effect	Deviance	*df*	AIC	comparison	X^2(*df*)	*p*
m0	-	2352	3	2358			
m1	bVt Odd, bVt ID	2313	5	2323	m0-m1	39.2(2)	<.001***
m2	Diverse Words Odd, Diverse Words ID	2306	5	2316	m0-m2	46.3(2)	<.001***
m3	**bVt Odd, bVt ID, Diverse Words Odd, Diverse Words ID**	**2301**	**7**	**2315**	**m1-m3**	**11.4(2)**	**<.01****
					m2-m3	4.3(2)	>.05
m4	bVt Odd, bVt ID, Diverse Words Odd, Diverse Words ID, Language group	2300	8	3516	m3-m4	1.4(1)	>.05

Note: Random effects included Participant and Item for all models. AIC = Akaike information criterion. LRT = Likelihood ratio test. Odd = oddity task. ID = identification task. Bold text reflects the best fit model.

/i, ɪ/ Correct Responses with Diverse Sentences

	score		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	4.74	3.61 – 6.22	**<0.001**
c_i_odd_bvt	1.00	0.99 – 1.01	0.745
c_i_ID_bvt	1.01	1.00 – 1.03	0.129
c_i_ID_dw	1.03	1.01 – 1.05	**0.009**
c_i_odd_dw	1.02	1.00 – 1.04	0.035
Language [2]	0.81	0.57 – 1.14	0.226
Random Effects			
σ^2	3.29		
$\tau_{00\ \text{item}}$	0.84		
$\tau_{00\ \text{participant}}$	0.03		
ICC	0.21		

N participant	38
N item	64
Observations	2432
Marginal R^2 / Conditional R^2	0.085 / 0.277

Note: for /i, ɪ/, results largely mirror results with the control. Language no longer becomes a significant predictor as control has been removed and Mandarin and Korean were not significantly different.

Experiment 1 /ɛ, æ/ Model comparison of prompt and language predictors for Diverse Sentences. Table shows model with all data (top) and an adjusted model with overlapping words removed (bottom).

Model	Fixed effect	Deviance	*df*	AIC	LRT comparison	*X*^{2}(*df*)	*p*
m0	-	2454	3	2460			
m1	**bVt Odd, bVt ID**	**2419**	**5**	**2429**	**m0-m1**	**35.3(2)**	**<.001*****
m2	**Diverse Words Odd, Diverse Words ID**	**2419**	**5**	**2429**	**m0-m2**	**35.4(2)**	**<.001*****
m3	bVt Odd, bVt ID, Diverse Words Odd, Diverse Words ID	2415	7	2429	m1-m3	4.4(2)	>.05
					m2-m3	4.3(2)	>.05
m4	bVt Odd, bVt ID, Diverse Words Odd, Diverse Words ID, Language group	2415	8	2431	m3-m4	0.03(1)	>.05

Note: Random effects included Participant and Item for all models. AIC = Akaike information criterion. LRT = Likelihood ratio test. Odd = oddity task. ID = identification task. Bold text reflects the best fit model.

Note: when control and language is omitted for /ɛ, æ/, bVt and Diverse Words perform nearly identically (O*R* = 1.04 for both Diverse Words ID and bVt ID, the same OR as /i, ɪ/ without language).

	score		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	2.64	2.07 – 3.37	**<0.001**
c_e_odd_bvt	1.00	0.99 – 1.01	0.817

c_e_ID_bvt	1.02	1.00 – 1.05	0.065
c_e_ID_dw	1.02	1.00 – 1.05	0.092
c_e_odd_dw	1.00	0.99 – 1.02	0.620
language [2]	1.05	0.61 – 1.81	0.861
Random Effects			
σ^2	3.29		
$\tau_{00\ \mathrm{item}}$	0.33		
$\tau_{00\ \mathrm{participant}}$	0.13		
ICC	0.12		
$N_{\mathrm{participant}}$	38		
N_{item}	56		
Observations	2128		
Marginal R^2 / Conditional R^2	0.097 / 0.207		

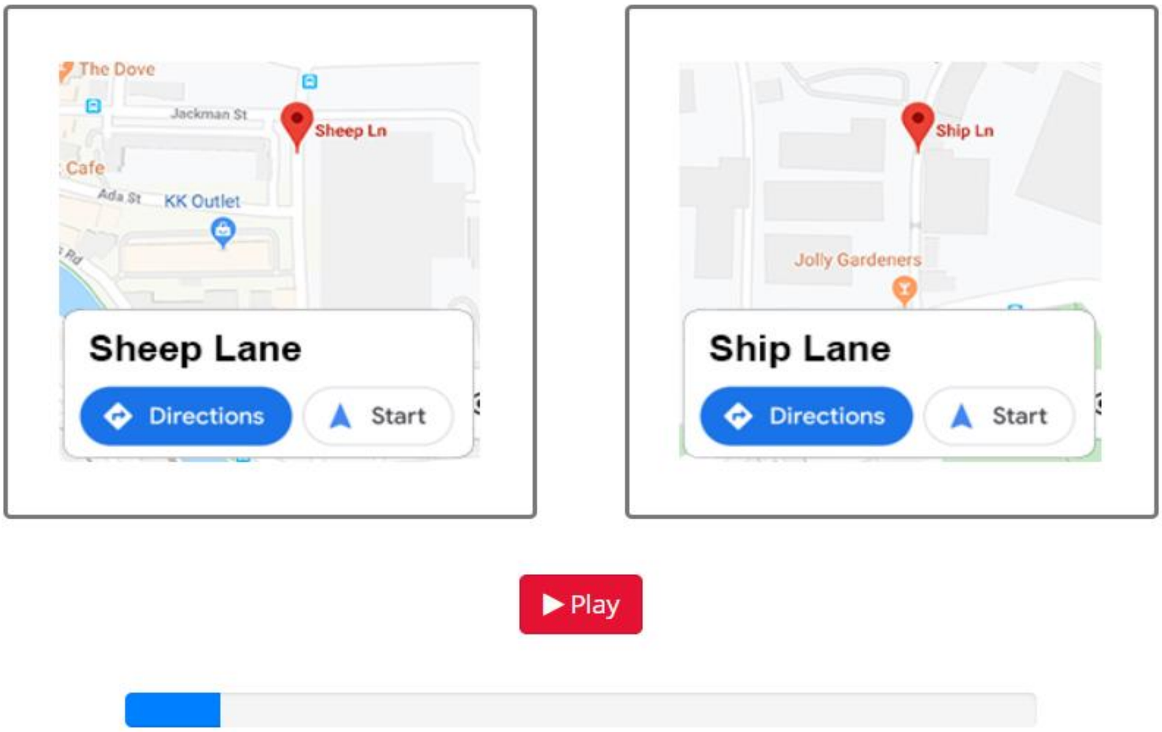
Note: for /ɛ, æ/, when language is included and the control is excluded, no predictors are significant.

Experiment 1 /i, ɪ/ Adjusted model comparison for generalised linear mixed models. Overlap between Diverse Words and Sentences removed.

Model	Fixed effect	Deviance	*df*	AIC	LRT comparison	X^2(*df*)	*p*
m0	-	1845	3	1851			
m1	bVt Odd	1819	4	1827	m0-m1	26.7(1)	<.001
m2	bVt ID	1814	4	1822	m0-m2	30.9(1)	<.001
m3	Diverse Words Odd	1806	4	1814	m0-m3	39.2(1)	<.001
m4	**Diverse Words ID**	1796	4	1804	m0-m4	49.7(1)	<.001
m5	Directions	1811	4	1819	m0-m5	34.3(1)	
m6	Language	1835	4	1843	m0-m6	44.9(1)	<.001

Note: Random effects included Participant and Item for all models. AIC = Akaike information criterion. LRT = Likelihood ratio test. Bold text reflects the best fit model.

Directions identification task screenshot



Experiment 1 Open-ended prompt

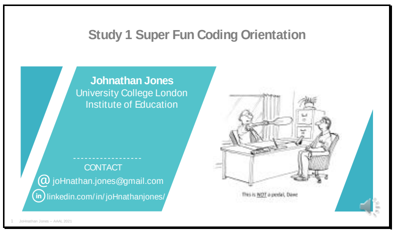
Well done!

We'd like to hear how you felt about the different parts of this experiment, and welcome comments below. To refer to individual parts of the experiment, here's what you did (not necessarily in order):

1. **b-vowel-t discrimination** (e.g., choose the odd word in "bet-bat-bet-bet")
2. **diverse words discrimination** (i.e., choosing the odd word in words other than b-vowel-t, such as "mill-meal-mill-mill")
3. **b-vowel-t identification** (e.g., selecting buttons labelled "bit" or "beat")
4. **diverse words identification** (i.e., selecting buttons labelled with words other than b-vowel-t, such as "lid" or "lead")
5. **diverse sentence identification** (i.e., selecting buttons labelled with words heard in a sentence, such as "the Dutch have basic mills")
6. **controlled sentence identification** (i.e., selecting buttons labelled with street names heard in sentences, as in "Meet me at Ship Lane")

Experiment 1 Coder Training

Slide 1



Study 1 Super Fun Coding Orientation

Johnathan Jones
University College London
Institute of Education

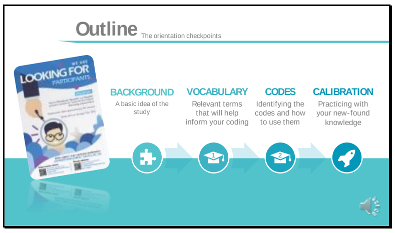
CONTACT
@ joHnathan.jones@gmail.com
in linkedin.com/in/joHnathanjones/

This is NOT a pedal, Dave

1 JoHnathan Jones – AAAL 2021

Hi, welcome to your coding orientation! I'm **Johnathan** and today we'll be going over some things to ensure we're on the same page for coding. This is for a study which examined how well second language learners were able to perceive the difference between English vowel sounds.

Slide 2



Outline The orientation checkpoints

WE ARE LOOKING FOR PARTICIPANTS

BACKGROUND A basic idea of the study

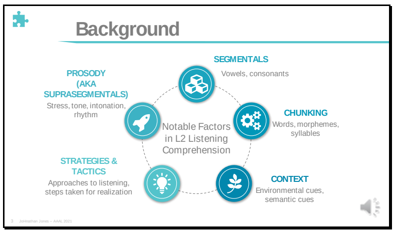
VOCABULARY Relevant terms that will help inform your coding

CODES Identifying the codes and how to use them

CALIBRATION Practicing with your new-found knowledge

Briefly, I'll be giving you some background on what you're looking at, then some terms that you'll need to understand to sort the data. We'll of course look at the codes, and then we'll finish off with some calibration. All fun stuff. Ready? Let's go!

Slide 3



Background

Notable Factors in L2 Listening Comprehension

SEGMENTALS Vowels, consonants

CHUNKING Words, morphemes, syllables

CONTEXT Environmental cues, semantic cues

STRATEGIES & TACTICS Approaches to listening, steps taken for realization

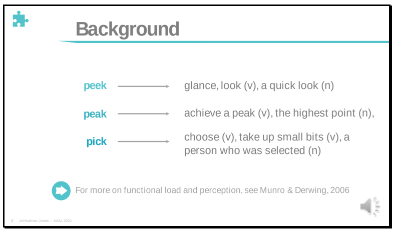
PROSODY (AKA SUPRASEGMENTALS) Stress, tone, intonation, rhythm

3 JoHnathan Jones – AAAL 2021

So this study was interested in assessing listening. There are many important considerations for listening perception, but I focused on segmentals, and more specifically, vowels. But what makes them so important?
They have a high functional load. If you have trouble differentiating certain vowels, as many second language

speakers do, things can get a little more complicated than they would be otherwise.

Slide 4



Background

peek	glance, look (v), a quick look (n)
peak	achieve a peak (v), the highest point (n),
pick	choose (v), take up small bits (v), a person who was selected (n)

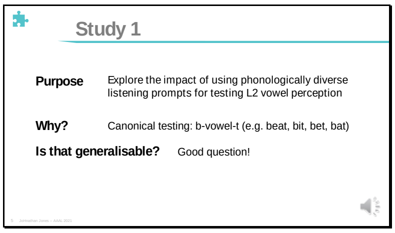
For more on functional load and perception, see Munro & Derwing, 2006

4 JoHnathan Jones – AAAL 2021

Think of homophones in your first language, words that sound the same but represent different things. Like **peak**, P-E-A-K, and **peek**, P-E-E-K. Now if you have a hard time differentiating between the high /i/ and the slightly lower /I/, you’ve **added** to your list of homophonous words, so you hear P-I-C-K the same as peak and peek. So there is a compound effect in play, and **compounding** is a feature we find in sounds that have a high functional load.

Of course, context and grammar help disambiguate things. But it isn’t always so simple. This research looks at cases where things aren’t so simple, and there’s an intersection between accuracy and intelligibility. In other words, you need to hear things accurately in order to correctly perceive the speakers intent.

Slide 5



Study 1

Purpose Explore the impact of using phonologically diverse listening prompts for testing L2 vowel perception

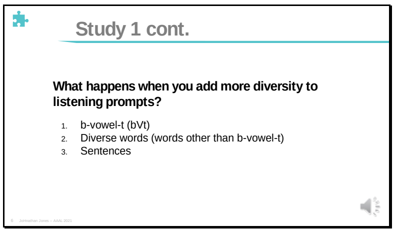
Why? Canonical testing: b-vowel-t (e.g. beat, bit, bet, bat)

Is that generalisable? Good question!

5 JoHnathan Jones – AAAL 2021

The purpose of this study was to investigate the impact of using diverse listening prompts for testing L2 vowel perception. Vowel perception experiments usually use an isolated consonantal frame for testing vowel perception. It is often monosyllabic, like b-vowel-t. This is done because vowels are easily influenced by their neighbours. Everytime you change a vowel's neighbouring consonant, the vowel sound changes a bit. So bat and back, for example, have different AH sounds. And if you add in a sentence context, then you've got connected speech influencing things. Keeping the consonantal environment to a fixed frame keeps results nice and clean.

Slide 6



Study 1 cont.

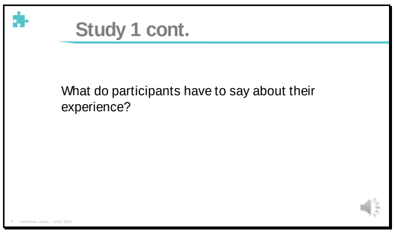
What happens when you add more diversity to listening prompts?

1. b-vowel-t (bVt)
2. Diverse words (words other than b-vowel-t)
3. Sentences

6 JoHnathan Jones – AAAL 2021

So this study asked, what happens if we use more diverse environments? Does a participant's ability to perceive the difference between vowel in a fixed, isolated frame like b-vowel-t translate to them being able to perceive the difference in more diverse environments? We looked at how well participants were able to perceive each vowel using different prompt types. We used the classic b-vowel-t, but we also used other prompt types, like Diverse Words and Sentences

Slide 7

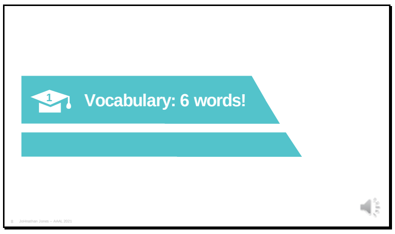


Study 1 cont.

What do participants have to say about their experience?

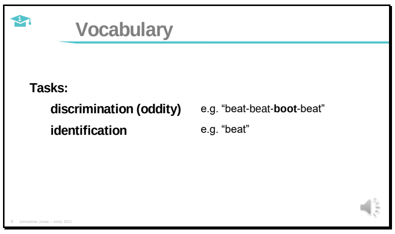
In addition to scores, we wanted to know what participant experience was like. So at the end of the experiment, we asked if there is anything they'd like to say about the experiment or its individual components. This is where you come in. Let's get you familiar with the vocabulary.

Slide 8



Vocabulary: 6 words!

Slide 9



Vocabulary

Tasks:

discrimination (oddity)	e.g. “beat-beat-**boot**-beat”
identification	e.g. “beat”

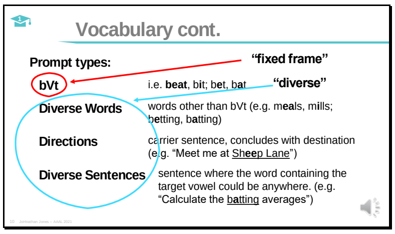
9 JoHnathan Jones – AAAL 2021

28 participants for Study 1 wrote comments for an open-ended question. Don’t worry, they didn’t write much—I promise!—but to understand the comments, you’ll need to know certain vocabulary. Participants used terms specific to the Study to explain their experiences.

So there were two task types in the study: discrimination and identification. For discrimination, four words were played in succession, and one word was a different word than the other three. For example, beat-beat-boot-beat. Here, the third word, boot, is the odd one out.

For identification, participants heard a single, isolated word, like “beat”. They were shown minimally paired words, and had to indicate which they heard.

Slide 10



Vocabulary cont.

Prompt types:

"fixed frame"

"diverse"

bVt — i.e. **beat**, b**i**t; b**e**t, b**a**t

Diverse Words — words other than bVt (e.g. m**ea**ls, m**i**lls; b**e**tting, b**a**tting)

Directions — carrier sentence, concludes with destination (e.g. "Meet me at Sh**ee**p Lane")

Diverse Sentences — sentence where the word containing the target vowel could be anywhere. (e.g. "Calculate the b**a**tting averages")

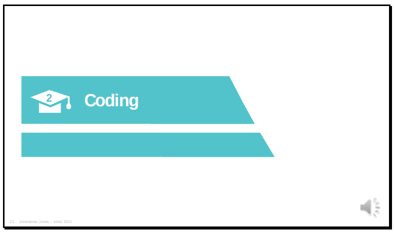
10 JoHnathan Jones – AAAL 2021

And for the last bit of vocabulary, there were four basic prompt types:

- B-vowel-t, as previously described.
- Diverse Words, which included words other than b-vowel-t
- Directions, a carrier sentence with a destination at the end of it.
- And finally, Diverse Sentences, where the target vowel was in a word that could be found anywhere in the sentence.

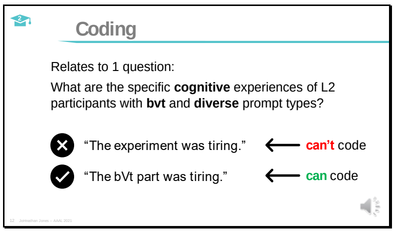
It's possible you'll see the terms "fixed frame" and "diverse". Or maybe I'll say it without realising it. Fixed frame simply means b-vowel-t; it's formulaic. Diverse means…well…not fixed.
Feel free to stop here and look at the examples or refer back to it as needed. Next we'll look at the codes.

Slide 11



Coding

Slide 12



Coding

Relates to 1 question:

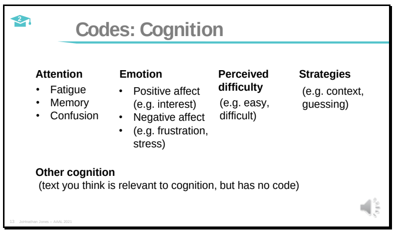
What are the specific **cognitive** experiences of L2 participants with **bvt** and **diverse** prompt types?

- ☒ “The experiment was tiring.” ⟵ **can't** code
- ✓ “The bVt part was tiring.” ⟵ **can** code

Remember when you’re doing this that we are trying to understand the differences between the classic bVt prompt and more diverse prompts when assessing vowel perception. This means that coding should be specific to a prompt type (bvt, Diverse Words, Directions, Diverse Sentences).

Saying the experiment was tiring is not useful. We can’t really code that for a specific prompt type. But saying bVt was tiring is useful. We’ll come back to this example after you have a look at the codes.

Slide 13



Codes: Cognition

Attention
- Fatigue
- Memory
- Confusion

Emotion
- Positive affect (e.g. interest)
- Negative affect
- (e.g. frustration, stress)

Perceived difficulty
(e.g. easy, difficult)

Strategies
(e.g. context, guessing)

Other cognition
(text you think is relevant to cognition, but has no code)

13 JoHnathan Jones – AAAL 2021

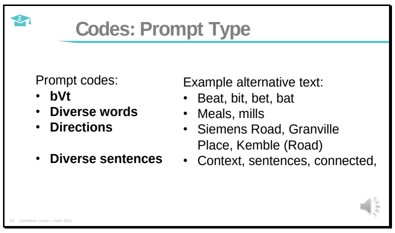
Because we’re interested in cognition related to prompt types, there are two main things you’ll need to code for: **cognition** and **prompt type**. First we’ll look at cognition.

Cognition has four main codes: attention, emotion, perceived difficulty, and strategies. Attention and emotion have additional subcodes, as seen here. Attention includes fatigue, memory, and confusion. Emotion has positive and negative affect. Positive affect would be something like interest, while negative affect would be frustration or stress.

Perceived difficulty and strategies don’t have additional codes, but we can clarify what to expect. Perceived difficulty is any reference to difficulty, whether easy or hard.

It’s possible you may find something you think is relevant to cognition, but is not included here. For that, we have a fifth category: Other cognition. You may never need this category, but it’s there if needed.

Slide 14

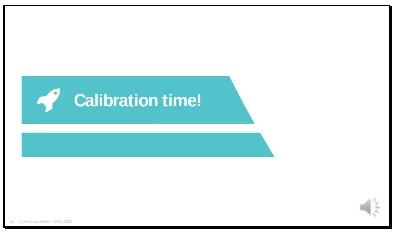


Codes: Prompt Type

Prompt codes:	Example alternative text:
• **bVt**	• Beat, bit, bet, bat
• **Diverse words**	• Meals, mills
• **Directions**	• Siemens Road, Granville Place, Kemble (Road)
• **Diverse sentences**	• Context, sentences, connected,

So that was Cognition, let's go over the codes for prompt type. These are the same 4 you saw in the vocabulary section. The only thing to note here is that sometimes participants road specific word pairs instead of the prompt. For Diverse words, you may see something like meals/mills, while for directions you'll see Siemens Road, Granville Place, or Campbell Road. If you encounter something you're not sure of, let me know!

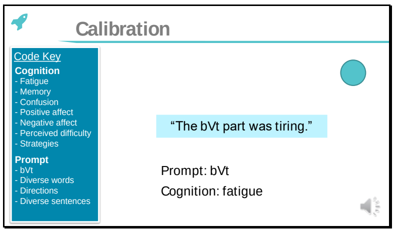
Slide 15



Calibration time!

For calibration, we'll be going over some selected texts together, and then I'll leave you with a self assessment component.

Slide 16



Calibration

Code Key

Cognition
- Fatigue
- Memory
- Confusion
- Positive affect
- Negative affect
- Perceived difficulty
- Strategies

Prompt
- bVt
- Diverse words
- Directions
- Diverse sentences

"The bVt part was tiring."

Prompt: bVt

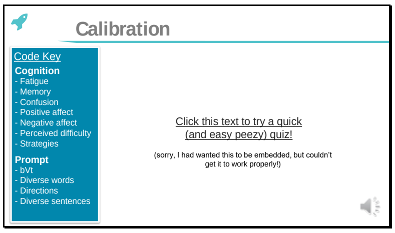
Cognition: fatigue

Now that we have our codes, let's return to our first coding example. With the Code Key on the left, How should we code this text? Take a moment to think about it. Just blurt it out if you have the answer. I'll put a timer in the upper right.

Did you get it? The first thing we need to do when looking at the participant responses is see if the text passes our criteria. Remember, we have two main criteria: prompt type and cognition. Does this text pass that test? Yes. (click) BVT tells us they are referring to the bVt prompt. And (click) tiring suggests fatigue.

And we're going to code the whole (click) sentence twice. (click) Once for bVt as a prompt and (click) once for fatigue as a type of cognition.

Slide 17



Calibration

Code Key

Cognition
- Fatigue
- Memory
- Confusion
- Positive affect
- Negative affect
- Perceived difficulty
- Strategies

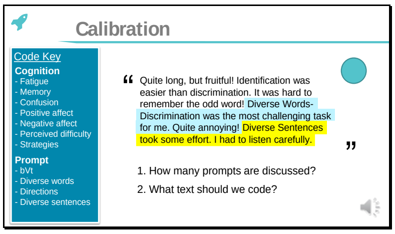
Prompt
- bVt
- Diverse words
- Directions
- Diverse sentences

Click this text to try a quick (and easy peezy) quiz!

(sorry, I had wanted this to be embedded, but couldn't get it to work properly!)

If you click the link, it will take you to an online quiz. When done with the online quiz, come back here and continue to the next slide.

Slide 18



Calibration

Code Key

Cognition
- Fatigue
- Memory
- Confusion
- Positive affect
- Negative affect
- Perceived difficulty
- Strategies

Prompt
- bVt
- Diverse words
- Directions
- Diverse sentences

“Quite long, but fruitful! Identification was easier than discrimination. It was hard to remember the odd word! Diverse Words- Discrimination was the most challenging task for me. Quite annoying! Diverse Sentences took some effort. I had to listen carefully.”

1. How many prompts are discussed?
2. What text should we code?

Welcome back. So we should have a working understanding of the codes at this point and how to apply them. Let's try something with a little more context.

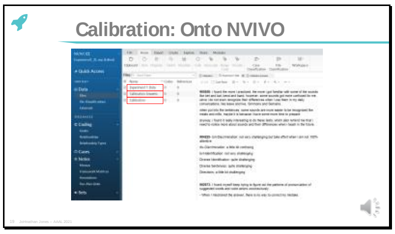
Take a minute to read the response on your screen. (click) How many prompts does it discuss? (click) What text should we code?

After break: ok, so hopefully that was enough time. If not, we can do it together. How many prompts were discussed?

We see diverse (click) words and Diverse (click) sentences. So two. What text should we code? Looking back to Diverse Words, it reads (click, click)…"challenging" refers to difficulty. And "quite annoying" suggests negative affect. We can code the whole thing three times. Once for the prompt type, and twice for cognition. How about Diverse Sentences? (click) Diverse Sentences took some effort. I had to listen carefully. This is perceived difficulty. So this text would be double coded.

Why didn't we touch the text at the beginning? It doesn't reference a specific prompt. Now one more slide and we're done.

Slide 19



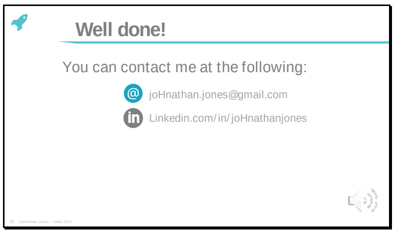
Calibration: Onto NVIVO

Ok, so we’re almost done. For the last little bit, you’ll need to open up the NVIVO coding file.

There, you’ll see three files: Experiment 1 Data, calibration answers, and calibration.

This gives you a chance to try coding with real data. This practice calibration data has been excluded from the main data set. The codes can be accessed using the codes panel on the left. The version of NVIVO you see here is NVIVO 20. This uses codes instead of nodes. If you have any compatibility issues, let me know and I’ll see if I can copy to a legacy version.

Slide 20



Well done!

You can contact me at the following:

@ joHnathan.jones@gmail.com

in Linkedin.com/in/joHnathanjones

20 JoHnathan Jones – AAAL 2021

Ok, so we’re almost done. For the last little bit, you’ll need to open up the NVIVO coding file.

There, you’ll see three files: Experiment 1 Data, calibration answers, and calibration.

This gives you a chance to try coding with real data. This practice calibration data has been excluded from the main data set. The codes can be accessed using the codes panel on the left. The version of NVIVO you see here is NVIVO 20. This uses codes instead of nodes. If you have any compatibility issues, let me know and I’ll see if I can copy to a legacy version.

Slide 21

Well done, now get to it!!

If you have any questions, contact me
at the following emails:
joHnathan.jones@gmail.com
joHnathan.jones.17@ucl.ac.uk

Table of results PAM study of Mandarin L1

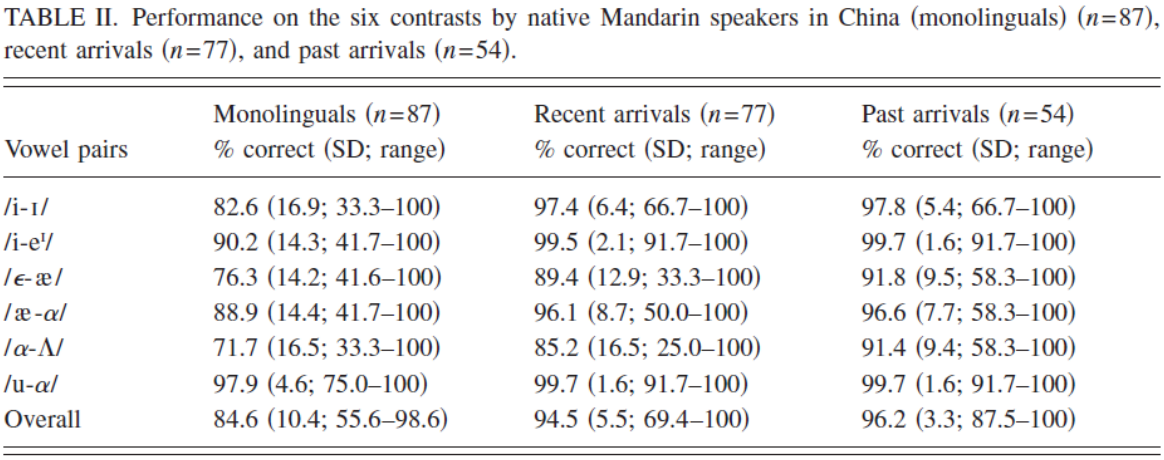


TABLE II. Performance on the six contrasts by native Mandarin speakers in China (monolinguals) (*n*=87), recent arrivals (*n*=77), and past arrivals (*n*=54).

Vowel pairs	Monolinguals (*n*=87) % correct (SD; range)	Recent arrivals (*n*=77) % correct (SD; range)	Past arrivals (*n*=54) % correct (SD; range)
/i-ɪ/	82.6 (16.9; 33.3–100)	97.4 (6.4; 66.7–100)	97.8 (5.4; 66.7–100)
/i-e^{ɪ}/	90.2 (14.3; 41.7–100)	99.5 (2.1; 91.7–100)	99.7 (1.6; 91.7–100)
/ϵ- æ/	76.3 (14.2; 41.6–100)	89.4 (12.9; 33.3–100)	91.8 (9.5; 58.3–100)
/ æ -α/	88.9 (14.4; 41.7–100)	96.1 (8.7; 50.0–100)	96.6 (7.7; 58.3–100)
/α-Λ/	71.7 (16.5; 33.3–100)	85.2 (16.5; 25.0–100)	91.4 (9.4; 58.3–100)
/u-α/	97.9 (4.6; 75.0–100)	99.7 (1.6; 91.7–100)	99.7 (1.6; 91.7–100)
Overall	84.6 (10.4; 55.6–98.6)	94.5 (5.5; 69.4–100)	96.2 (3.3; 87.5–100)

Experiment 2 Appendix

Experiment 1 Corrected point biserial index for /ɛ, æ/ (bVt Identification with Diverse Sentence total score)

Participant	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	/ɛ, æ/ Dive
1	1	1	1	1	1	1	1	1	26
4	1	1	1	1	1	1	1	1	38
5	1	1	1	1	1	1	1	0	24
6	1	1	1	1	1	1	1	1	28
7	1	1	1	1	1	1	1	1	25
8	1	1	1	1	1	1	1	1	28
9	1	1	1	1	1	1	1	1	35
10	1	1	1	0	1	1	1	1	27
11	0	0	1	0	1	1	1	1	25
12	1	1	1	1	1	1	0	1	32
13	1	1	1	1	1	1	1	0	19
14	1	1	1	1	1	1	1	1	36
15	1	0	1	1	1	1	1	1	30
16	1	1	1	1	1	1	1	1	32
19	1	1	1	1	1	1	1	1	23
20	1	1	1	1	1	1	1	1	26
21	0	0	0	0	1	1	1	1	23
22	0	1	1	0	1	1	1	0	26
23	1	1	1	1	1	1	1	1	44
24	1	0	0	0	1	1	1	1	35
25	1	1	1	1	1	1	1	0	33
26	1	1	1	1	1	1	1	1	28
27	1	1	1	1	1	1	1	1	27
28	1	1	1	1	1	1	1	1	46
31	1	1	1	0	0	1	1	1	25
32	1	1	1	0	1	1	0	1	24
33	0	0	1	0	1	1	1	1	39
35	1	1	1	1	1	1	1	1	44
36	0	0	0	0	1	1	1	1	22
38	1	0	1	0	0	0	0	0	22
39	1	1	1	1	1	1	1	0	27
40	1	1	1	1	1	1	1	1	30
41	1	1	1	1	1	1	1	1	33
43	0	1	1	1	0	1	1	1	25
44	0	0	0	0	1	1	1	1	24
45	1	1	1	1	1	1	1	1	26
46	1	1	1	1	1	1	1	1	21
48	1	1	1	1	1	1	1	1	25
49	1	1	1	1	1	1	1	1	44
51	1	1	1	1	1	1	1	1	27
52	1	1	1	1	1	1	1	0	33
53	1	1	1	1	1	1	0	0	40
Total	35	34	38	31	39	41	38	34	
Df	0.357143	0.346939	0.387755	0.316327	0.397959	0.418367	0.387755	0.346939	
Dc	0.221823	0.154792	0.174432	0.272925	0.229922	0.174971	0.009003	0.119459	
							Mean DC	0.169666	

Experiment 1 Corrected point biserial index for /i, ɪ/ (bVt Identification with Diverse Sentence total score)

Participant	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	/i, ɪ/ Divers
1	1	1	0	1	1	1	1	1	23
4	1	1	1	1	1	1	0	1	45
5	1	1	1	1	1	1	1	1	29
6	1	1	1	1	1	1	1	1	26
7	0	0	0	1	1	1	1	1	36
8	1	1	1	1	1	1	1	1	27
9	1	1	1	1	1	1	1	1	26
10	0	0	0	0	1	1	1	1	39
11	1	1	1	1	1	1	1	1	35
12	1	1	1	1	1	1	1	1	28
13	0	0	1	1	1	1	0	1	18
14	1	1	1	1	1	1	1	1	28
15	1	1	1	1	1	1	1	1	45
16	1	1	1	1	1	1	1	1	42
19	1	1	1	1	1	0	1	0	24
20	1	1	1	1	1	1	1	1	23
21	1	1	1	1	1	1	1	0	17
22	1	1	0	1	1	1	1	1	34
23	1	1	1	1	1	1	1	1	43
24	1	1	1	1	1	1	1	1	22
25	1	1	0	1	1	1	1	1	28
26	0	0	0	0	1	1	1	1	39
27	1	1	1	1	1	1	1	1	27
28	1	1	1	1	1	1	1	1	44
31	1	1	1	1	1	1	1	1	26
32	1	1	0	1	1	1	1	1	21
33	1	1	1	1	1	1	1	1	24
35	1	1	1	1	1	1	0	1	34
36	1	1	1	1	1	1	1	1	34
38	1	1	0	1	1	1	1	1	23
39	1	1	1	1	1	1	1	1	23
40	1	1	1	1	1	1	1	1	27
41	1	1	1	1	1	1	0	1	28
43	1	1	1	1	1	1	0	1	24
44	1	0	0	1	1	1	1	1	23
45	1	1	1	1	1	1	1	1	29
46	0	0	0	0	1	1	1	1	32
48	1	0	1	1	1	1	1	1	26
49	1	1	1	1	1	1	1	1	43
51	1	1	1	1	1	1	1	1	30
52	0	1	0	1	1	1	1	1	41
53	1	1	1	1	1	1	1	0	33
Total	36	35	31	39	42	41	37	39	
Df	0.367347	0.357143	0.316327	0.397959	0.428571	0.418367	0.377551	0.397959	
Dc	-0.21122	-0.01254	-0.04709	-0.23426	0	0.127042	0.019936	0.20141	
							Mean Dc	-0.01959	

Experiment 2 Recruitment poster (English)

WE ARE
LOOKING FOR
PARTICIPANTS

VOLUNTEER!

Native **Mandarin**, **Spanish**, and **English** speakers needed for online British English **listening experiment**

Experiment takes approximately 90 minutes

Study will run through May 2020

EMAIL SUBJECT LINE: "LISTENING EXPERIMENT"
JOHNATHAN.JONES.17@UCL.AC.UK

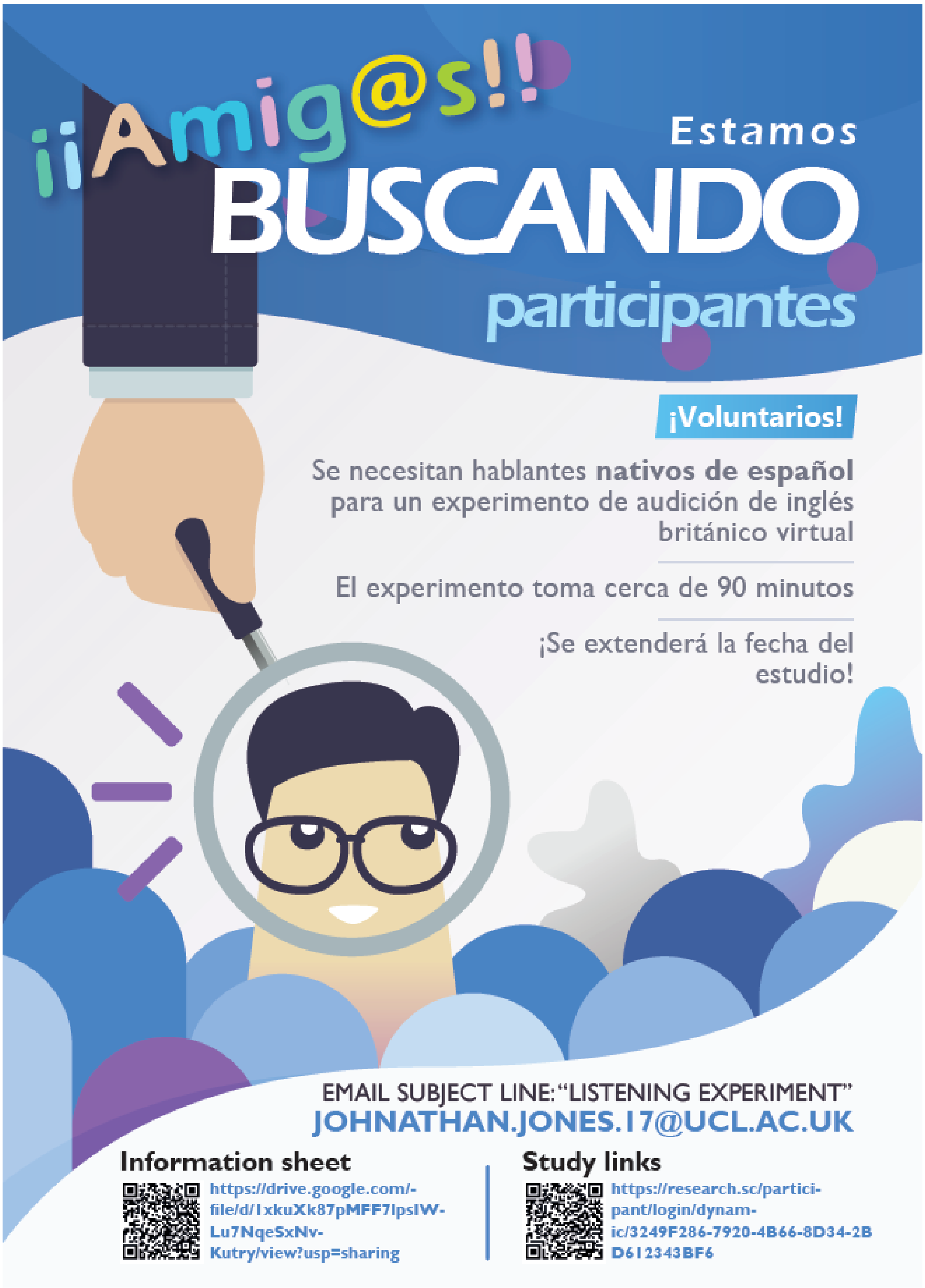
Information sheet

https://drive.google.com/-
file/d/1H5Rs4lATta8xOW19RuL-
rYHD1-CnM-
Hobv/view?usp=sharing

Study address

https://research.sc/partici-
pant/login/dynam-
ic/3249F286-7920-4B66-8D34-2B
D612343BF6

Experiment 2 Recruitment poster (Spanish)



¡¡Amig@s!!

Estamos BUSCANDO participantes

¡Voluntarios!

Se necesitan hablantes **nativos de español** para un experimento de audición de inglés británico virtual

El experimento toma cerca de 90 minutos

¡Se extenderá la fecha del estudio!

EMAIL SUBJECT LINE: "LISTENING EXPERIMENT"

JOHNATHAN.JONES.17@UCL.AC.UK

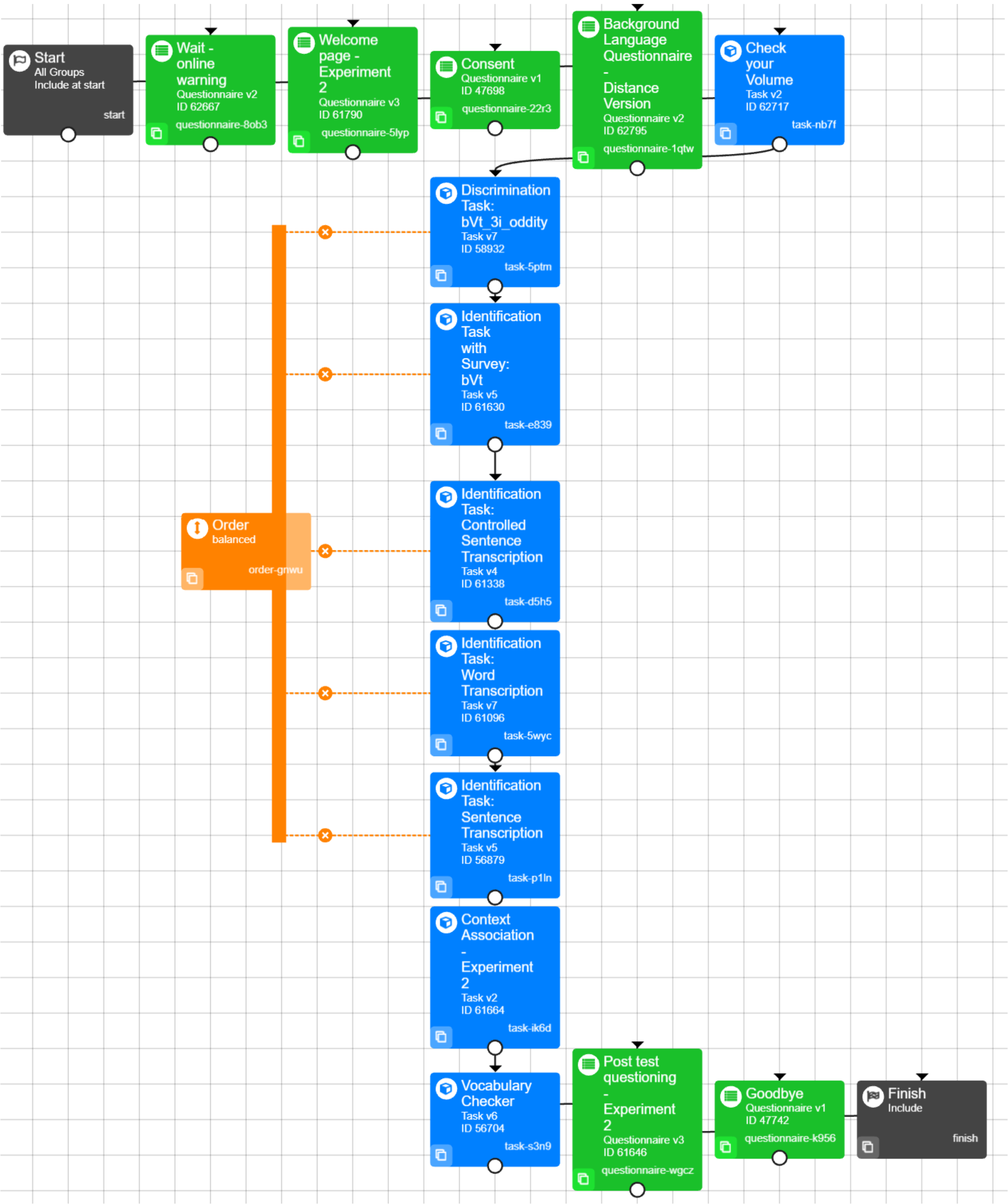
Information sheet

https://drive.google.com/-
file/d/1xkuXk87pMFF7lpslW-
Lu7NqeSxNv-
Kutry/view?usp=sharing

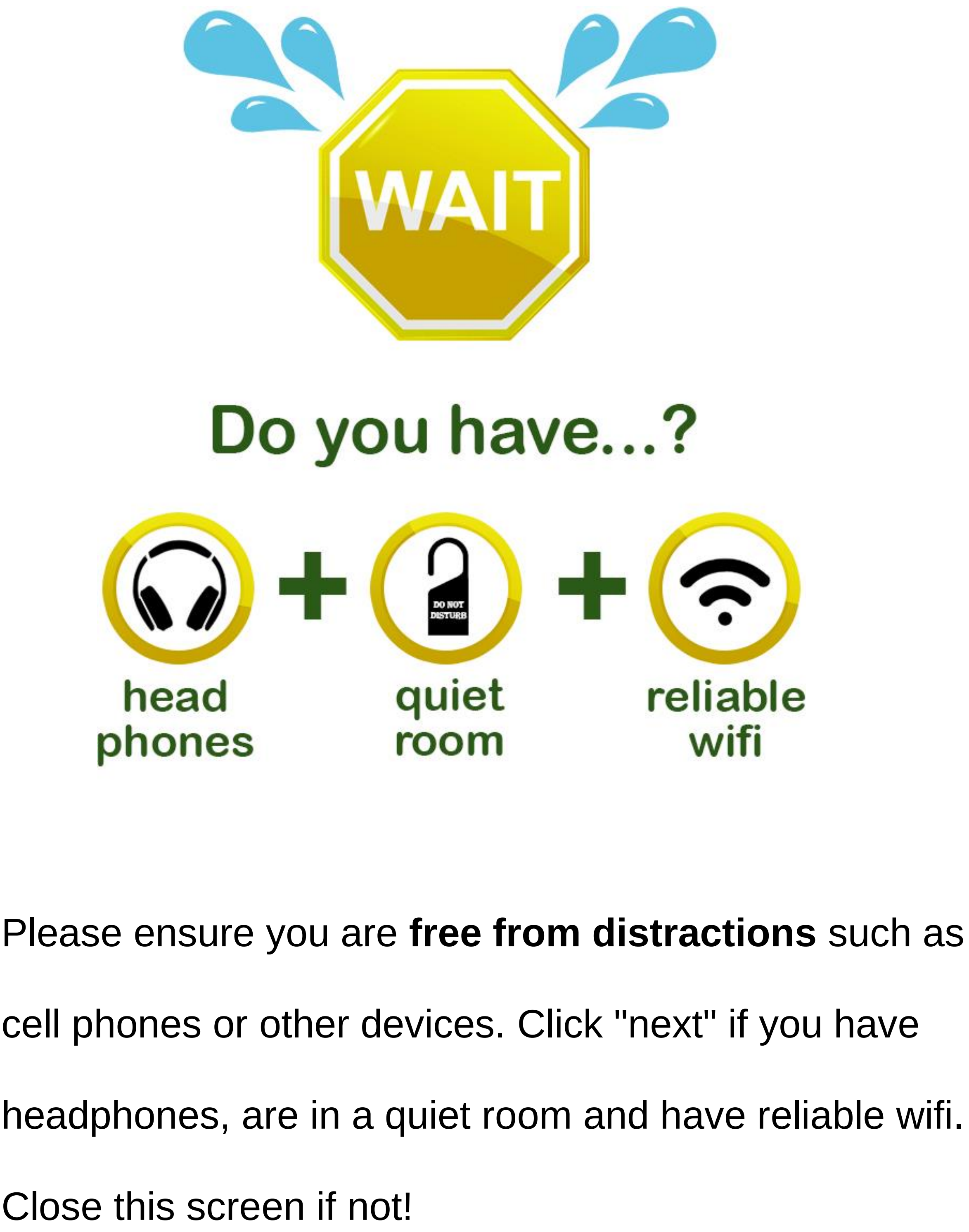
Study links

https://research.sc/partici-
pant/login/dynam-
ic/3249F286-7920-4B66-8D34-2B
D612343BF6

Experiment 2 design screenshot

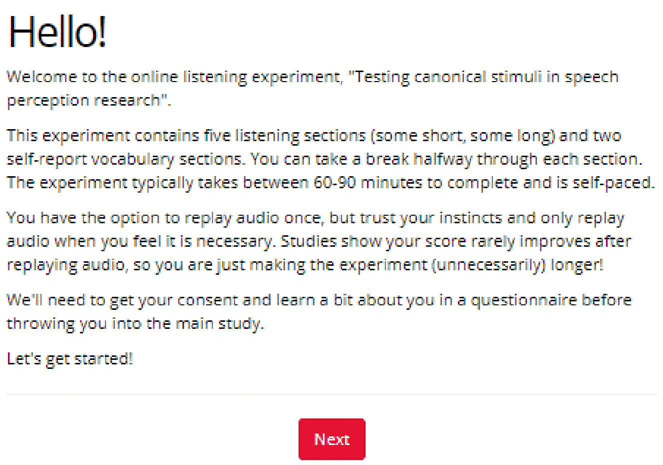


Warning screen screenshot



Please ensure you are **free from distractions** such as cell phones or other devices. Click "next" if you have headphones, are in a quiet room and have reliable wifi. Close this screen if not!

Welcome screen screenshot



Hello!

Welcome to the online listening experiment, "Testing canonical stimuli in speech perception research".

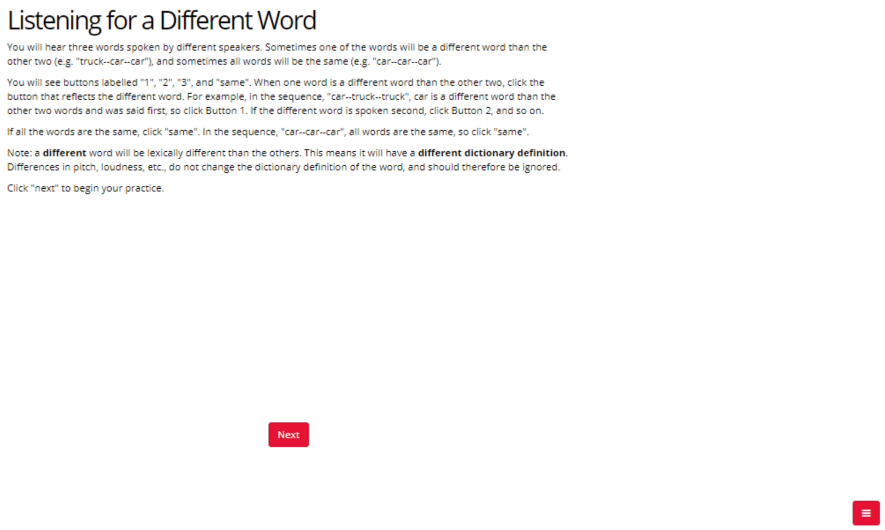
This experiment contains five listening sections (some short, some long) and two self-report vocabulary sections. You can take a break halfway through each section. The experiment typically takes between 60-90 minutes to complete and is self-paced.

You have the option to replay audio once, but trust your instincts and only replay audio when you feel it is necessary. Studies show your score rarely improves after replaying audio, so you are just making the experiment (unnecessarily) longer!

We'll need to get your consent and learn a bit about you in a questionnaire before throwing you into the main study.

Let's get started!

Instructions screenshot



Listening for a Different Word

You will hear three words spoken by different speakers. Sometimes one of the words will be a different word than the other two (e.g. "truck--car--car"), and sometimes all words will be the same (e.g. "car--car--car").

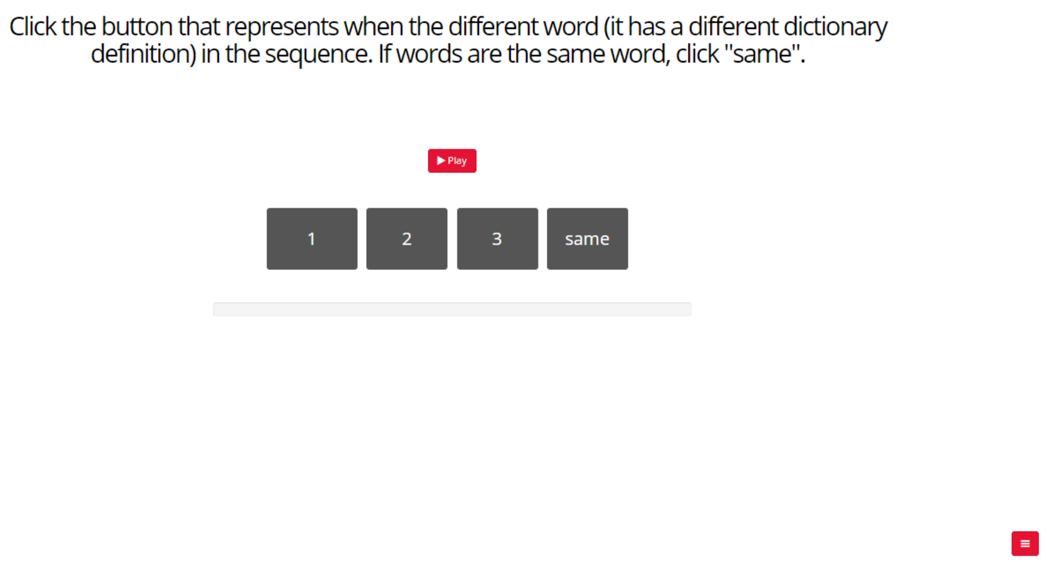
You will see buttons labelled "1", "2", "3", and "same". When one word is a different word than the other two, click the button that reflects the different word. For example, in the sequence, "car--truck--truck", car is a different word than the other two words and was said first, so click Button 1. If the different word is spoken second, click Button 2, and so on.

If all the words are the same, click "same". In the sequence, "car--car--car", all words are the same, so click "same".

Note: a **different** word will be lexically different than the others. This means it will have a **different dictionary definition**. Differences in pitch, loudness, etc., do not change the dictionary definition of the word, and should therefore be ignored.

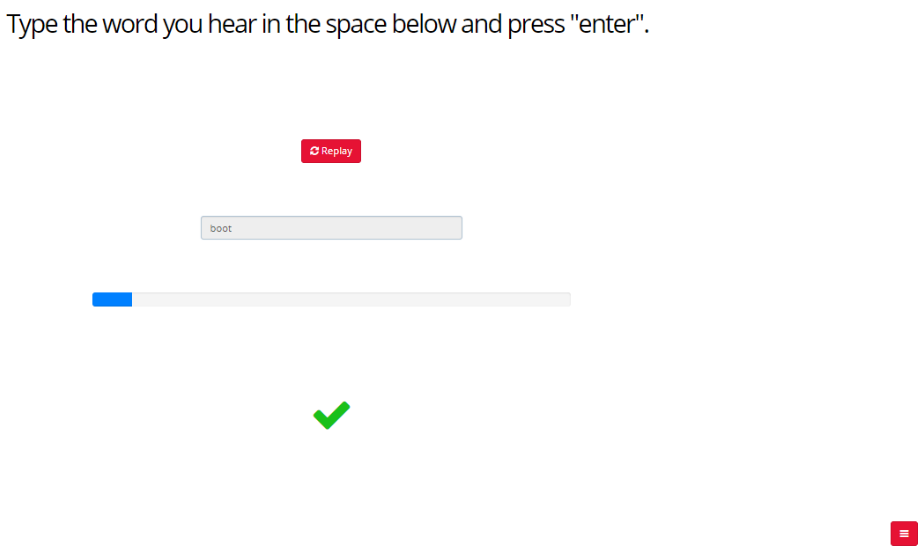
Click "next" to begin your practice.

Oddity task screenshot



Click the button that represents when the different word (it has a different dictionary definition) in the sequence. If words are the same word, click "same".

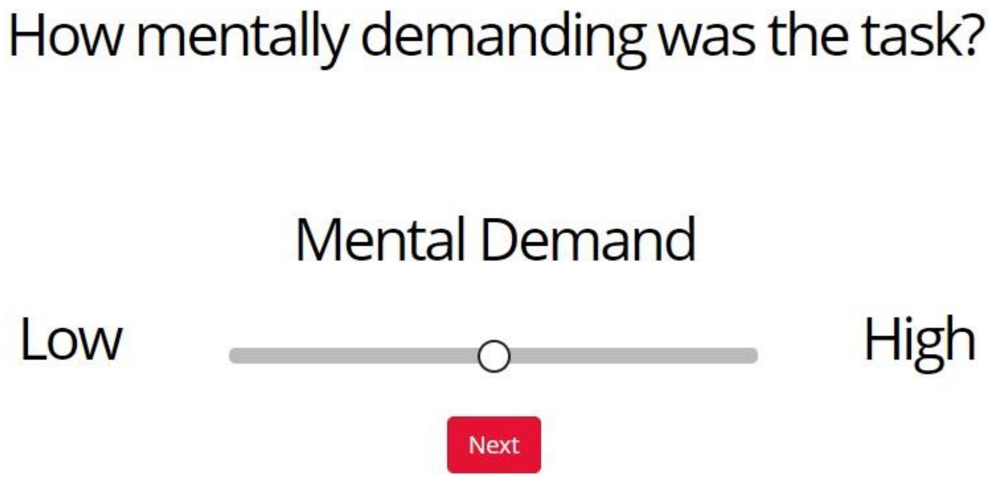
bVt Transcription task screenshot



Type the word you hear in the space below and press "enter".

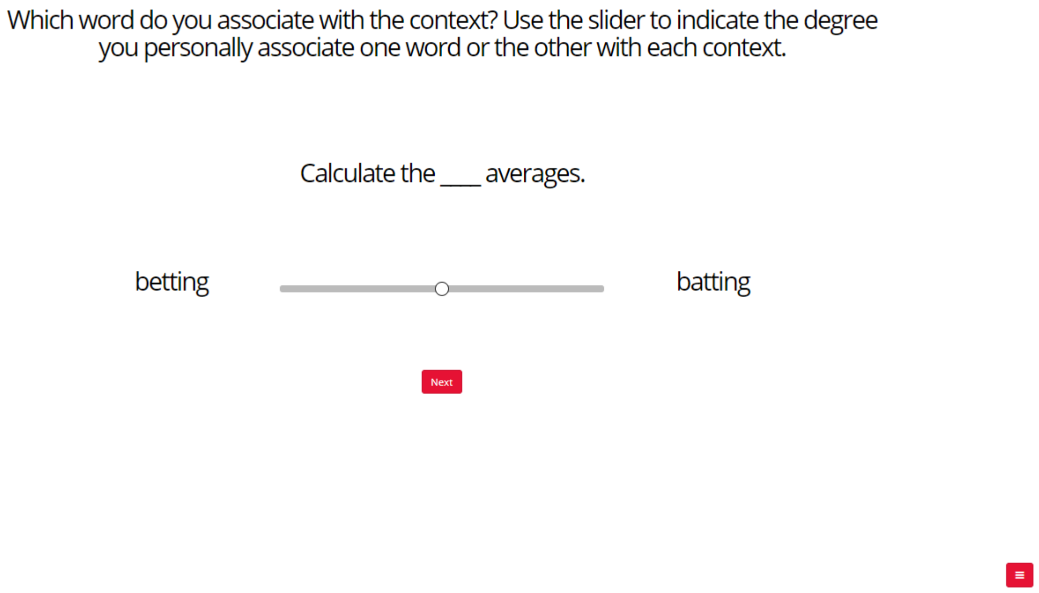
boot

NASA TLX screenshot



How mentally demanding was the task?

Association task screen shot

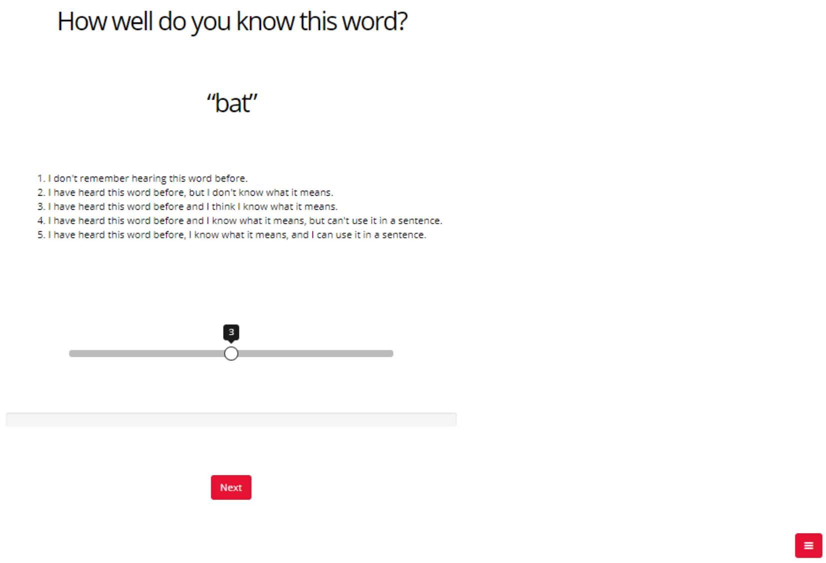


Which word do you associate with the context? Use the slider to indicate the degree you personally associate one word or the other with each context.

Calculate the ____ averages.

betting ——○—— batting

Target word familiarity survey screen shot



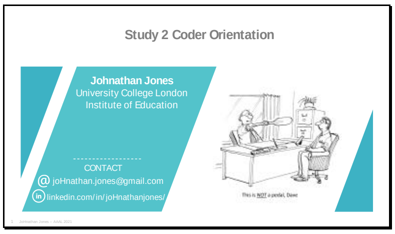
How well do you know this word?

“bat”

1. I don't remember hearing this word before.
2. I have heard this word before, but I don't know what it means.
3. I have heard this word before and I think I know what it means.
4. I have heard this word before and I know what it means, but can't use it in a sentence.
5. I have heard this word before, I know what it means, and I can use it in a sentence.

Experiment 2 Coder Training

Slide 1



Study 2 Coder Orientation

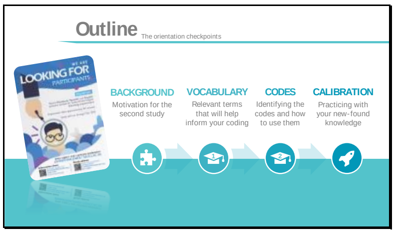
Johnathan Jones
University College London
Institute of Education

CONTACT
@ joHnathan.jones@gmail.com
in linkedin.com/in/joHnathanjones/

1 JoHnathan Jones – AAAL 2021

Hi, welcome to your coding orientation! I'm **Johnathan** and today we'll be going over some things to ensure we're on the same page for coding. This is for a study which examined how well second language learners were able to perceive the difference between English vowel sounds.

Slide 2



Outline The orientation checkpoints

BACKGROUND
Motivation for the second study

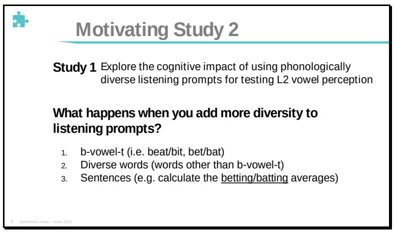
VOCABULARY
Relevant terms that will help inform your coding

CODES
Identifying the codes and how to use them

CALIBRATION
Practicing with your new-found knowledge

Briefly, I'll be giving you some background on what you're looking at, then some terms that you'll need to understand to sort the data. We'll of course look at the codes, and then we'll finish off with some calibration. All fun stuff. Ready? Let's go!

Slide 3



Motivating Study 2

Study 1 Explore the cognitive impact of using phonologically diverse listening prompts for testing L2 vowel perception

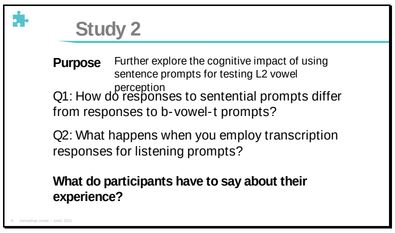
What happens when you add more diversity to listening prompts?

1. b-vowel-t (i.e. beat/bit, bet/bat)
2. Diverse words (words other than b-vowel-t)
3. Sentences (e.g. calculate the betting/batting averages)

3 JoHnathan Jones – AAAL 2021

Recall that traditional studies have typically used fixed consonantal frames, like b-vowel-t, for testing L2 vowel perception. Study 1 explored the use of more phonologically diverse environments for listening prompts. The study used diverse words and sentences, but the scores were potentially inflated by using a closed set of labelled options. Participants knew what the target words were because of the option labels, and thus could ignore the sentence and listen strictly for the target word. In the example sentence shown here (CLICK), participants would be able to listen strictly for either betting or batting. They wouldn’t necessarily have to understand the sentence. Further, they had a 50-50 chance of getting the item correct. So the question was, what would happen if we removed the cues and made responses an open set?

Slide 4



Study 2

Purpose Further explore the cognitive impact of using sentence prompts for testing L2 vowel perception

Q1: How do responses to sentential prompts differ from responses to b-vowel-t prompts?

Q2: What happens when you employ transcription responses for listening prompts?

What do participants have to say about their experience?

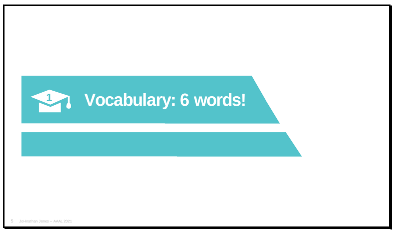
4 JoHnathan Jones – AAAL 2021

To answer this, we decided to try transcription tasks in a second study. We knew that there could be spelling confounds, but felt that this could be mitigated through the use of high-intermediate-to-advanced adult speakers. Participants were told to do their best with spelling, but to spell the word how it sounds to them if they weren't sure. The study was interested in how participants perceived target vowels, so prescriptive spelling was irrelevant so long as participants were able to communicate what they heard.

As with Study 1, we wanted to know not only how participants would perform, but what their experience was like. So at the end of the experiment, we asked if there were anything they'd like to say about the experiment or its individual components. This is where you come in. You'll be coding what participants had to say about their experience.

For Study 2, vocabulary and coding is similar to that of Study 1, but with a few additions. Let's go over the relevant vocabulary and coding.

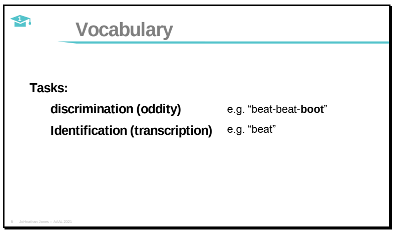
Slide 5



Vocabulary: 6 words!

5 JoHnathan Jones – AAAL 2021

Slide 6



Vocabulary

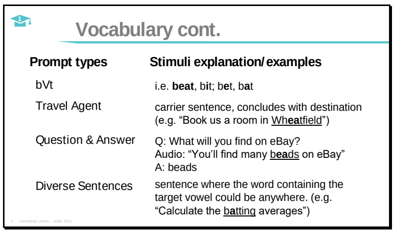
Tasks:

discrimination (oddity)	e.g. “beat-beat-**boot**”
Identification (transcription)	e.g. “beat”

6 JoHnathan Jones – AAAL 2021

Recall that Study 1 used discrimination (CLICK) and identification (CLICK) tasks. For the discrimination task, participants had to listen for the odd word out (CLICK), while for the identification task, participants heard a single word and had to identify which word they heard. Study 2 used these tasks as well. There were some slight tweaks, but those aren’t relevant here, so we’ll skip the details.

Slide 7



Vocabulary cont.

Prompt types	Stimuli explanation/ examples
bVt	i.e. **beat**, b**i**t; b**e**t, b**a**t
Travel Agent	carrier sentence, concludes with destination (e.g. “Book us a room in Wh**ea**tfield”)
Question & Answer	Q: What will you find on eBay? Audio: “You’ll find many b**ea**ds on eBay” A: beads
Diverse Sentences	sentence where the word containing the target vowel could be anywhere. (e.g. “Calculate the b**a**tting averages”)

7 JoHnathan Jones – AAAL 2021

And for the last bit of vocabulary, there were four basic prompt types:

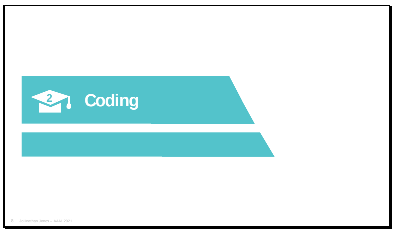
- B-vowel-t, which you’re quite familiar with.
- Travel agent, which is similar to the Direction Task in Study 1. There’s a carrier sentence and a destination at the end. Participants had to listen and write the name of the destination. These were all real places in the UK, but were generally unknown to participants.
- Question & Answer, Participants read a question that focuses their attention on what to listen

for. Then they heard an audio recording and wrote the answer based on what they heard.

- And finally, Diverse Sentences, where the target vowel was in a word that could be found anywhere in the sentence. Participants had to write the entire sentence and were not cued as to what information to listen for.

Feel free to stop here and look at the examples or refer back to this slide as needed. Next we'll look at the codes.

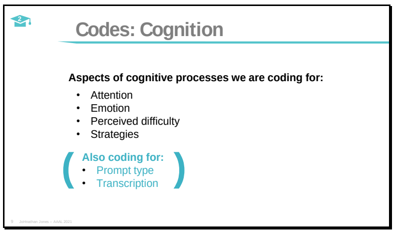
Slide 8



2 Coding

8 JoHnathan Jones – AAAL 2021

Slide 9



Codes: Cognition

Aspects of cognitive processes we are coding for:

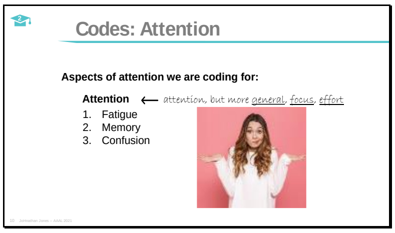
- Attention
- Emotion
- Perceived difficulty
- Strategies

Also coding for:

- Prompt type
- Transcription

We're interested in cognition related to prompt types, and we're also interested in the effects of transcription. First we'll look specifically at cognition.

Slide 10



Codes: Attention

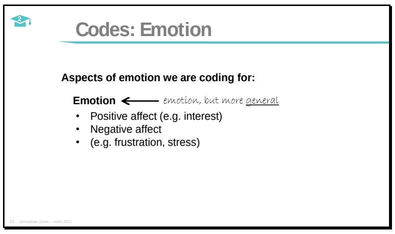
Aspects of attention we are coding for:

Attention ← attention, but more general, focus, effort

1. Fatigue
2. Memory
3. Confusion

Attention (CLICK) has three subcodes (CLICK): fatigue, memory, and confusion. (CLICK) What if the subcodes don't cover everything? If you find that the text you're looking at is related to attention, but doesn't fit the subcodes, you can use the parent code (CLICK), Attention. This is for more general elements of attention, and includes focus and effort.

Slide 11



Codes: Emotion

Aspects of emotion we are coding for:

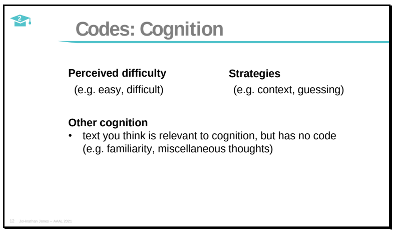
Emotion ← *emotion, but more general*

- Positive affect (e.g. interest)
- Negative affect
- (e.g. frustration, stress)

11 JoHnathan Jones – AAAL 2021

Emotion (CLICK) has two subcodes (CLICK): positive affect and negative affect. Positive affect would be something like interest, while negative affect would relate to frustration or stress. As with Attention, if you find that the text you're looking at is related to emotion, but doesn't fit the subcodes, you can use the parent code, Emotion (CLICK). It is possible you don't use this parent code, but it's there if needed.

Slide 12



Codes: Cognition

Perceived difficulty
(e.g. easy, difficult)

Strategies
(e.g. context, guessing)

Other cognition

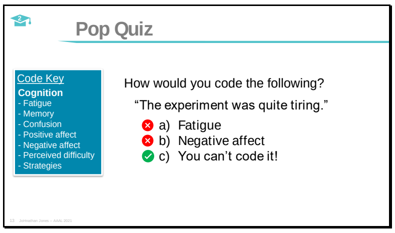
- text you think is relevant to cognition, but has no code (e.g. familiarity, miscellaneous thoughts)

12 JoHnathan Jones – AAAL 2021

Perceived difficulty and strategies don't have additional codes, but we can clarify what to expect. Perceived difficulty is any reference to difficulty, whether easy or hard. Strategies are anything participants use to answer the prompts that don't include actual listening perception. Maybe they use context to answer the question, or guessing.

It's possible you may find something you think is relevant to cognition, but is not included here. For that, we have a fifth category: Other cognition. Examples of Other Cognition include "familiarity" or "miscellaneous thoughts about specific prompts".

Slide 13



Pop Quiz

Code Key
Cognition
- Fatigue
- Memory
- Confusion
- Positive affect
- Negative affect
- Perceived difficulty
- Strategies

How would you code the following?

"The experiment was quite tiring."

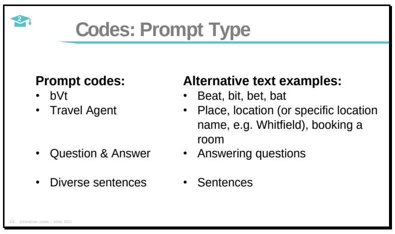
✗ a) Fatigue
✗ b) Negative affect
✓ c) You can't code it!

13 JoHnathan Jones – AAAL 2021

All right, now that we have some codes, let's do a pop quiz. How would you code the text, "The experiment was quite tiring?" I'll put some time on the clock.
Ok, done? Tiring refers to fatigue, but this is actually the wrong answer. The correct answer is 3, you can't code it. Why not? Because the text does not refer to a specific prompt. It doesn't help us answer our research question.

So now let's have a look at the prompt codes that we match the cognition codes with.

Slide 14

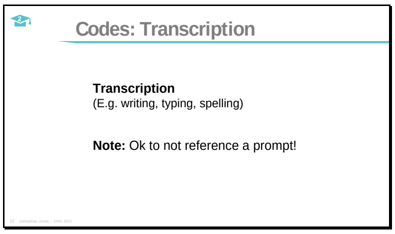


Codes: Prompt Type

Prompt codes:	Alternative text examples:
• bVt	• Beat, bit, bet, bat
• Travel Agent	• Place, location (or specific location name, e.g. Whitfield), booking a room
• Question & Answer	• Answering questions
• Diverse sentences	• Sentences

These are the same 4 you saw in the vocabulary section. The only thing to note here is that sometimes participants road specific word pairs instead of the prompt. For bVt, they could say the actual word; for Travel Agent, they might use the words place, or location synonymously with the task. Or they might just talk about booking a room. For question and answer, they may refer to answering questions, and for Diverse Sentences, they may simply say sentences. If you encounter something you're not sure of, let me know!

Slide 15



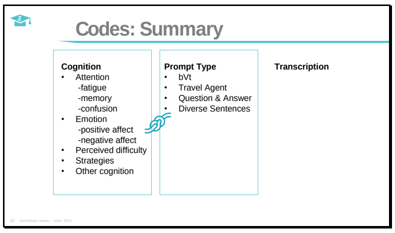
Codes: Transcription

Transcription
(E.g. writing, typing, spelling)

Note: Ok to not reference a prompt!

Finally, anything related to transcription will be lumped into the same code: transcription. When you see someone referring to writing, typing, or spelling, it goes here. (CLICK) Note that transcription is different than cognition for coding. While participants have to reference a prompt for you to code cognition, they don't for transcription. References to transcription can receive a code even if they do not refer to a prompt.

Slide 16



Codes: Summary

Cognition
- Attention
  - -fatigue
  - -memory
  - -confusion
- Emotion
  - -positive affect
  - -negative affect
- Perceived difficulty
- Strategies
- Other cognition

Prompt Type
- bVt
- Travel Agent
- Question & Answer
- Diverse Sentences

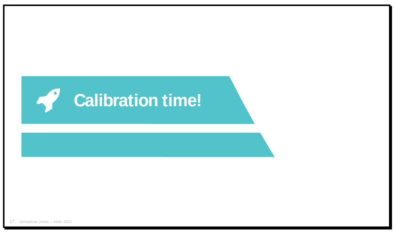
Transcription

16 JoHnathan Jones – AAAL 2021

To summarise, we are largely interested in the cognitive processes (CLICK) that participants discuss. How those relate to prompt type is essential (CLICK), so cognition must always refer to a prompt. Transcription is also important, and can stand on its own.

That's all the codes. Time for calibration!

Slide 17

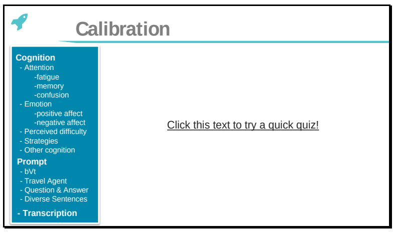


Calibration time!

17 JoHnathan Jones – AAAL 2021

For calibration, we'll be going over some selected texts together, and then I'll leave you with a self assessment component.

Slide 18



Calibration

Cognition
- Attention
  - -fatigue
  - -memory
  - -confusion
- Emotion
  - -positive affect
  - -negative affect
- Perceived difficulty
- Strategies
- Other cognition

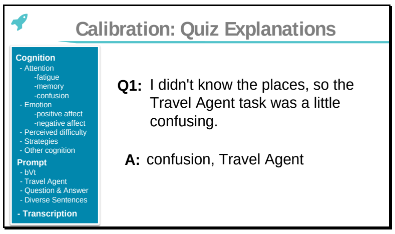
Prompt
- bVt
- Travel Agent
- Question & Answer
- Diverse Sentences

- Transcription

Click this text to try a quick quiz!

If you click the link, it will take you to an online quiz. When done with the online quiz, come back here and continue to the next slide.

Slide 19



Calibration: Quiz Explanations

Cognition
- Attention
  - -fatigue
  - -memory
  - -confusion
- Emotion
  - -positive affect
  - -negative affect
- Perceived difficulty
- Strategies
- Other cognition

Prompt
- bVt
- Travel Agent
- Question & Answer
- Diverse Sentences

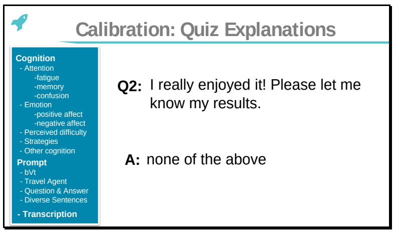
- Transcription

Q1: I didn't know the places, so the Travel Agent task was a little confusing.

A: confusion, Travel Agent

Quiz Exp
There are two required codes here: confusion and Travel Agent. Confusing indicates confusion, which is cognition. Cognition must link to a prompt type, so the prompt type, Travel Agent, must be indicated as well.

Slide 20



Calibration: Quiz Explanations

Cognition
- Attention
  - -fatigue
  - -memory
  - -confusion
- Emotion
  - -positive affect
  - -negative affect
- Perceived difficulty
- Strategies
- Other cognition

Prompt
- bVt
- Travel Agent
- Question & Answer
- Diverse Sentences

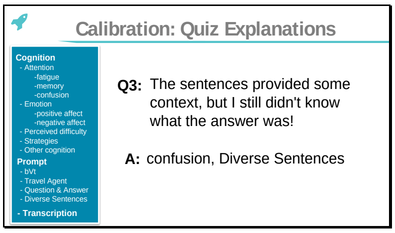
- Transcription

Q2: I really enjoyed it! Please let me know my results.

A: none of the above

The cognition is Positive Affect, but because no prompt type is given, it can't help us explain certain prompt types. If there is no prompt type with cognition, we can't code it.

Slide 21



Calibration: Quiz Explanations

Cognition
- Attention
  - -fatigue
  - -memory
  - -confusion
- Emotion
  - -positive affect
  - -negative affect
- Perceived difficulty
- Strategies
- Other cognition

Prompt
- bVt
- Travel Agent
- Question & Answer
- Diverse Sentences

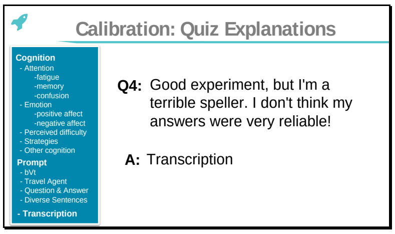
- Transcription

Q3: The sentences provided some context, but I still didn't know what the answer was!

A: confusion, Diverse Sentences

Not knowing or being unaware suggests confusion. Sentences is a synonym for Diverse Sentences.

Slide 22



Calibration: Quiz Explanations

Cognition
- Attention
  - -fatigue
  - -memory
  - -confusion
- Emotion
  - -positive affect
  - -negative affect
- Perceived difficulty
- Strategies
- Other cognition

Prompt
- bVt
- Travel Agent
- Question & Answer
- Diverse Sentences

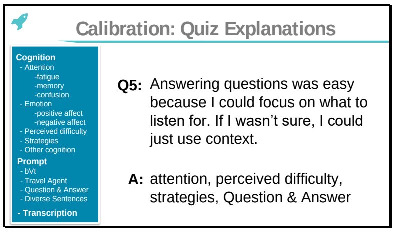
- Transcription

Q4: Good experiment, but I'm a terrible speller. I don't think my answers were very reliable!

A: Transcription

Only a single code can be applied here. Good experiment suggests positive affect, but there is no prompt type indicated. Terrible speller reflects transcription, and we can code this because unlike cognition with prompt types, transcription can stand on its own. Talking about reliability would be a good option for "other cognition" if a prompt were mentioned, but it was not linked with a prompt type. Instead, it describes the effects of transcription.

Slide 23



Calibration: Quiz Explanations

Cognition
- Attention
  - -fatigue
  - -memory
  - -confusion
- Emotion
  - -positive affect
  - -negative affect
- Perceived difficulty
- Strategies
- Other cognition

Prompt
- bVt
- Travel Agent
- Question & Answer
- Diverse Sentences

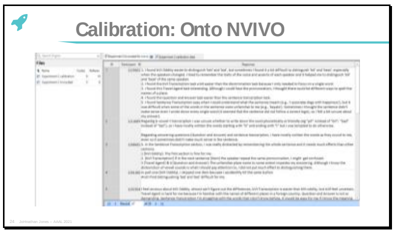
- Transcription

Q5: Answering questions was easy because I could focus on what to listen for. If I wasn't sure, I could just use context.

A: attention, perceived difficulty, strategies, Question & Answer

For the final Question, question 5, Question and Answer was the prompt, cued by the words "answering questions". The broad-category Attention was coded here to reflect focus needed for the question and answer task. The words, "if I wasn't sure, I could listen for context" is an example of strategies used by the participant. The person was using something other than the sounds of the words to decide what the answer was.
Now you're ready for your coding practice!

Slide 24



Calibration: Onto NVIVO

24 JoHnathan Jones – AAAL 2021

For the last little bit, you’ll need to open up the NVIVO coding file.

There, you’ll see two files: Experiment 2 data and Experiment 2 calibration.

This gives you a chance to try coding with real data. This practice calibration data is based on data which was excluded from the main data. Practice with the calibration data and give me a call! Leave the main Experiment 2 data alone for now.

Slide 25

Well done!

You can contact me at the following:

joHnathan.jones@gmail.com

Linkedin.com/in/joHnathanjones

Calibration notes

Study 2 Calibration Notes

Note: yellow = missing/not aligned coding; green = aligned coding. All misaligned coding is included below; aligned text is only included to highlight relevant coding differences and similarities.

Key areas of departure include:

1. coding both bVt tasks
2. overall coverage of coded text.
3. coding Other Cognition

Prompts

Sentence Transcription, bVt Oddity, Q&A, Travel Agent

bVt transcription

- Very confusing. Difficult to differenciate vowels, especially when the words were isolated.
- bVt was easiest, but boring.
    - bVt references (including "isolated words") were coded for bVt oddity, but not bVt transcription

Attention

- Sentences were really good, but took a lot of effort. I couldn't always remember the full sentence. The question task was good because I could focus on what I should listen for.
    - I included the subsequent yellow sentence to help give cause and effect context to the preceding text in green.

Memory

- Sentences were really good, but took a lot of effort. I couldn't always remember the full sentence.
    - I included the preceding yellow sentence to help give cause and effect context to the text in green.

Confusion

- Very confusing. Difficult to differenciate vowels, especially when the words were isolated.
    - Would you say "very confusing" was for the experiment in general? I can see that, but I linked it to bVt tasks as the participant clarified exacerbation when "words were isolated". I could go either way here. What do you think?

Fatigue

Affect (neutral)

Positive affect

- Sentences were really good, but took a lot of effort. I couldn't always remember the full sentence. The question task was good because I could focus on what I should listen for.

Negative affect

Other cognition

- Travel agent was most difficult to tell the difference. Especially because I didn't know the places.
  - When a participant discusses their familiarity with vocabulary or tasks, or explains their thought processes, I use "other cognition". The participant stating they "didn't know places" suggests (lack of) familiarity. The preceding sentence is included to show the context for the cause and effect relationship (i.e. unfamiliarity = difficulty)

Perceived difficulty

- Very confusing. Difficult to differenciate vowels, especially when the words were isolated.

Strategies

Transcription

Unsure how to code

Correlational matrices for association for bVt transcription and Diverse Sentences

Correlation matrices show how the levels of associations correlate across prompt types. The strongest correlations were with equal and opposite associations and Diverse Sentences.

/i, ɪ/ Correlation Matrix for association in bVt transcription and Diverse Sentences

		Same	Opposite	Equal	Oddity	bVt ID	Diverse Sentences
Same	*r*	1	-0.113	.504*	.299*	.322*	.552**
	p		0.454	0.012	0.043	0.029	<0.01
	n	46	46	24	46	46	46
Opposite	*r*	-0.113	1	0.197	.385**	.415**	.672**
	p	0.454		0.357	0.008	0.004	<0.01
	n	46	46	24	46	46	46
Equal	*r*	.504*	0.197	1	.678**	0.400	.755**
	p	0.012	0.357		0.000	0.053	<0.01
	n	24	24	24	24	24	24
Oddity	*r*	.299*	.385**	.678**	1	.574**	.606**
	p	0.043	0.008	<0.01		<0.01	<0.01
	n	46	46	24	46	46	46
bVt ID	*r*	.322*	.415**	0.400	.574**	1	.559**
	p	0.029	0.004	0.053	<0.01		<0.01
	n	46	46	24	46	46	46
Diverse Sentences	*r*	.552**	.672**	.755**	.606**	.559**	1
	p	<0.01	<0.01	<0.01	<0.01	<0.01	
	n	46	46	24	46	46	46

Note. *r* = Pearson *r*; *p* = statistical significance; *n* = sample size.

* $p < 0.05$ (2-tailed).

** $p < 0.01$ (2-tailed).

/ɛ, æ/ Correlation Matrix for association in bVt transcription and Diverse Sentences

		Same	Opposite	Equal	Oddity	bVt ID	Diverse Sentences
Same	*r*	1	0.171	0.294	.353*	.339*	.593**
	p		0.257	0.209	0.016	0.021	<0.01
	n	46	46	20	46	46	46
Opposite	*r*	0.171	1	.445*	.409**	0.287	.885**
	p	0.257		0.049	0.005	0.053	<0.01

	n	46	46	20	46	46	46
Equal	*r*	0.294	.445*	1	0.207	0.251	.572**
	p	0.209	0.049		0.381	0.285	0.008
	n	20	20	20	20	20	20
Oddity	*r*	.353*	.409**	0.207	1	.365*	.501**
	p	0.016	0.005	0.381		0.013	<0.01
	n	46	46	20	46	46	46
bVt ID	*r*	.339*	0.287	0.251	.365*	1	.426**
	p	0.021	0.053	0.285	0.013		0.003
	n	46	46	20	46	46	46
Diverse Sentences	*r*	.593**	.885**	.572**	.501**	.426**	1
	p	<0.01	<0.01	0.008	<0.01	0.003	
	n	46	46	20	46	46	46

Note. *r* = Pearson *r*; *p* = statistical significance; *n* = sample size

* *p* < 0.05 (2-tailed).

** *p* < 0.01 (2-tailed).

Correlations between IELTS scores and self-reported proficiency levels

		IELTS Overall	IELTS Listening	Self-report Overall	Self-report Listening	Overall Score
IELTS Overall	Correlation Coefficient	1.000	.822**	.134	.137	.076
	Sig. (2-tailed)	.	.000	.375	.365	.618
	N	46	46	46	46	46
IELTS Listening	Correlation Coefficient	.822**	1.000	.132	.213	.098
	Sig. (2-tailed)	.000	.	.383	.156	.519
	N	46	46	46	46	46
Self-report Overall	Correlation Coefficient	.134	.132	1.000	.590**	.209
	Sig. (2-tailed)	.375	.383	.	.000	.164
	N	46	46	46	46	46
Self-report Listening	Correlation Coefficient	.137	.213	.590**	1.000	.191
	Sig. (2-tailed)	.365	.156	.000	.	.204
	N	46	46	46	46	46
Overall Score	Correlation Coefficient	.076	.098	.209	.191	1.000
	Sig. (2-tailed)	.618	.519	.164	.204	.
	N	46	46	46	46	46