Evaluating the scoring system of an AI-integrated app to assess foreign language phonological decoding

Research Methods in Applied Linguistics 4 (2025) 100257

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Evaluating the scoring system of an AI-integrated app to assess foreign language phonological decoding

James Turner [a,*], Alison Porter [a], Suzanne Graham [b], Travis Ralph-Donaldson [a], Heike Krüsemann [a], Pengchong Zhang [b], Kate Borthwick [a]

[a] *Department of Languages, Cultures and Linguistics, University of Southampton, United Kingdom*
[b] *Institute of Education, University of Reading, United Kingdom*

ARTICLE INFO

Keywords:
FL phonological decoding
Language assessment
Artificial intelligence
Young learners
FL reading
FL pronunciation

ABSTRACT

Phonological decoding in a foreign language (FL)—a two-part process involving first the ability to map written symbols to their corresponding sounds and second to pronounce them intelligibly—is foundational for reading and vocabulary acquisition. Yet assessing this skill efficiently and at scale in young learners remains a persistent challenge. Here, we introduce and evaluate the accuracy and effectiveness of a novel method for assessing FL phonological decoding using an AI-driven app that automatically scores children's pronunciation of symbol-sound correspondences. In a study involving 254 learners of French and Spanish (aged 10–11) across five UK primary schools, pupils completed a read-aloud task (14 symbol-sound correspondences) that was scored by the app's automatic speech recognition (ASR) technology. The validity of these automated scores was tested by fitting them as independent variables in regression models predicting human auditory coding. The multiple significant relationships between automated and human scores that were established indicate that there is great potential for ASR-based tools to reliably assess phonological decoding in this population. These findings provide the first large-scale empirical validation of an AI-based assessment of FL decoding in children, opening new possibilities, applicable to a range of languages being learnt, for scalable and efficient assessment.

1. Introduction

Foreign language (FL) phonological decoding is an essential underpinning for several aspects of language learning, including reading and vocabulary development (Woore, 2022). Assessing how well learners can decode is therefore an important part of understanding their overall linguistic development. In an alphabetic language, phonological decoding refers, on the one hand, to converting written text into its speech form (Nassaji, 2014) and is facilitated by establishing links between individual letters or combinations of letters ("symbols") and their pronunciations. For example, if learners of English become aware that the symbol <ph> is often pronounced [f], decoding new words containing these letters, e.g. *photo*, can be accomplished more easily and quickly. On the other hand, the ability to orally produce an appropriate and intelligible version of that sound (pronunciation) should also be viewed as a central aspect of phonological decoding. That is especially true for low-input instructed L2 contexts where it cannot be assumed that learners will encounter distinctive FL sounds (Young-Sholten & Piske, 2008), particularly if their teacher's language proficiency level and knowledge of phonology are weak (Collen & Duff, 2024).

[*] Corresponding author.
E-mail address: James.Turner@soton.ac.uk (J. Turner).

https://doi.org/10.1016/j.rmal.2025.100257
Received 30 September 2024; Received in revised form 21 August 2025; Accepted 22 August 2025
Available online 1 September 2025

Assessing phonological decoding poses a methodological challenge especially in young learner (YL) contexts (Clark, 2017; Darnell et al., 2017) and this task is complicated further in the FL classroom. Both formative and summative types of assessment in the classroom often consist of teacher-led oral tasks (Graham, 2020), but these have limited scalability and are reliant on subjective human judgement. For instance, assessors may be influenced by their first (L1) sound system, they may struggle with the cognitive pressure of processing and assessing speech in real time, and they may not have sufficient time to spend coding spoken material on an individual basis (see Section 2.4). This may be exacerbated in YL contexts where progress can be very slow and almost imperceptible (Porter, 2020).

There are also valid questions pertaining to the assessment itself: conventional research methods of assessing phonological decoding often involve read-aloud tasks, but these are labour-intensive because learners must be recorded and scored by assessors individually (Woore, 2022), and the coding scheme used (e.g. accurate vs. inaccurate) may overlook the nuanced variation which is especially apparent in YLs' developmental speech patterns (Cable et al., 2010; Porter, 2020). Alternative instruments include sound-alike tasks (Erler & Macaro, 2011; Woore et al., 2018), but previous research with YLs has not found learners' responses to such tasks to yield a sufficiently sensitive phonological decoding assessment scale (see e.g. Woore et al., 2018). We return to these challenges in Section 2.4.

Despite the obvious importance of research in this area, relatively few studies have focused on methods for assessing phonological decoding in FL contexts. Still fewer have considered the value of Artificial Intelligence (AI) in solving the methodological challenges outlined above. Given the rapid expansion of AI, it seems timely to investigate how reliably it can do so. The present study therefore sought to contribute new knowledge by exploring the extent to which an app that uses a Deep Neural Network (DNN) to assess FL phonological decoding offers a valid alternative to traditional methods of assessment. The app used for this project was created and licensed by Niter LTD, and a bespoke version of the app was developed in collaboration with University of Southampton and University of Reading to facilitate this research.[1] The objective of this study was to provide a preliminary validation of four different scores provided by the app: two that analyse the individual sounds produced for specific FL spellings that are embedded within words (sound-level scores) and two scores which evaluate the word produced as a whole (one word score focusing on only the sounds recognised, and one "overall" word score focusing on the sounds recognised in addition to other aspects of pronunciation such as variations in pitch, timing, amplitude and word stress). The in-built app scores were all calculated via strategic comparisons with native speaker exemplar recordings to create similarity scores (see Section 3.4). The sound scores generated by the app for 14 symbol-sound correspondences (7 in two different FLs) were compared against the values obtained from auditory coding by a trained phonetician (the first author). The app's scores at word-level were also compared to the author's auditory coding for comprehensibility (i.e. how easy it is to understand the word as a whole). Another author performed the same process of auditory coding for 10 % of the data and both sets of auditory coding (sound-level and word-level) were subject to stringent inter-rater reliability checks (reported in Section 3.5). We conclude by focusing on the methodological implications of these results as well as their implications for YL classrooms.

2. Background

2.1. Defining FL phonological decoding

Learning a word in an FL involves storing its meaning in long-term memory along with its graphological (written) form and phonological (sound) form (Erler & Macaro, 2011). The process of actively converting letters or "symbols" (Woore, 2021) into their speech sounds is known as "phonological decoding" and is an essential skill for both word learning and word retrieval (Sparks, 2021). Acquiring symbol-sound correspondences (SSCs) is not only integral to phonological decoding but also to successfully processing aural input, that is, recognising sounds and matching them to phonological (and graphological) mental representations in order to access meaning (Bassetti, 2024). The term SSCs, rather than GPCs (grapheme-phoneme correspondences), is used in the present study in accordance with Woore (2021) who makes the point that links can be made between units larger than graphemes and combinations of sounds/phonemes, such as "spelling bodies" (e.g. *ight* in *fight*). This is especially the case in languages such as English with "deep" orthographies, i.e. where symbols less consistently correspond to the same sounds (see Section 2.3). Although previous research that focuses on FL phonological decoding marks a distinction between decoding (recognising a sound in text form) and pronunciation (producing an appropriate sound for that symbol) (e.g. Hamada & Koda, 2008; Woore, 2021), the present study treats sound articulation as an integral component of the FL phonological decoding construct for two reasons. First, in order to analyse language learners' knowledge of SSCs in isolation and remove the confound of pronunciation difficulties, only SSCs can be analysed in read-aloud tasks for which there is no influence from the phonetics and phonology of the first language, or SSCs for which that influence is easy to disregard (e.g. Woore, 2021). This results either in using less sensitive test instruments such as sound-alike or rhyme tasks (see Section 2.4) or dramatically reducing the number of SSCs that can be analysed in a read-aloud task, potentially resulting in the systematic exclusion of large sets of SSCs that are likely harder to learn because they involve internalising new articulatory patterns (Bassetti, 2024). Secondly, if pronunciation of FL sounds is featured in the assessment of phonological decoding, YLs may be less likely to form mental representations of words with incorrect phonological specifications, thus preventing subsequent difficulties in speech

[1] Niter LTD was a non-academic partner for the current research project, but the French and Spanish content of the tool were developed within the research project rather than as part of the commercial product. Given that this tool was only a prototype for trial purposes, and that all aspects of the research design and analysis were independently conducted by the researchers at the aforementioned institutions, we do not believe the integrity and impartiality of the findings to have been compromised.

production and perception that research has shown to be pervasive (see e.g. Bassetti, 2024; Bassetti et al., 2015). A scale that combines both SSC understanding and an approximation of phonological knowledge may also be more strongly indicative of other linguistic skills such as speaking fluency, aural word recognition and foreign language processing, but this will need to be tested empirically in future research.[2]

2.2. The importance of assessing FL phonological decoding

Receiving feedback about linguistic performance is an important element for enabling YLs to progress along their language learning trajectories (Courtney & Graham, 2019). For instructors, obtaining a measure of performance in a specific task is insightful, especially when these results are likely to predict students' future outcomes. For example, because of a posited relationship between FL phonological decoding and literacy skills (Hamada & Koda, 2008; Jeon & Yamashita, 2014), assessing FL phonological decoding may help to identify which learners are likely to grow into expert readers and those who may be in need of greater support (Sparks, 2021). This is further motivated by the Simple View of Reading, *SVR*, (Gough & Tunmer, 1986), which characterises the reading process as a strategic combination of linguistic comprehension and phonological decoding. Although originally applied to first language reading, the *SVR* has also been used to successfully identify strong and weak readers in FL contexts (see e.g. Sparks, 2021 for an overview of the evidence from a variety of FLs).

In psycholinguistics, it has been argued that phonological decoding is especially important for reading less familiar or unknown words but for more familiar words, semantic information may be accessed directly from the orthography without needing to access phonological information. There is hence a dual method for accessing lexical meaning (Coltheart, 2005; Coltheart et al., 1993) and this likely extends to FL reading (Woore, 2021). However, unlike first language reading contexts, for a number of foreign learners, the written form of language may be the primary source of input (Bassetti, 2024; Young-Sholten & Piske, 2008), which could favour mappings being established directly between the written words and their meanings (i.e. the non-phonological route). Equally, it may result in the meaning of FL words being mapped to L1 sound representations due to already established L1 SSCs (Bassetti, 2024; Hayes-Harb & Barrios, 2021). In short, secure FL phonological decoding skills are essential either for committing new words to memory with target-like phonological representations, or updating words already learnt with more target-like sound representations so as to avoid a negative impact on FL perception and production (Bassetti, 2024; Bassetti et al., 2015). We return to important distinctions between L1 and FL decoding in Section 2.3, but the rationale for an effective phonological decoding strategy and a method for assessing improvement is evident.

From a methodological perspective, assessing phonological decoding accurately and at scale has the capacity to advance both first language and FL reading research. For example, it would allow L1 and FL research to reliably address timely questions such as whether phonics instruction in classroom settings is more likely to improve decoding skills and hence, it is assumed, reading outcomes, over other methods (Castles et al., 2018; Fletcher et al., 2021). Further, it would provide insight into whether the effects of phonics instruction on reading comprehension differ between L1 and FL contexts, highlighting the need for an effective FL phonological decoding assessment tool all the more. Several FL YL studies have shown only a limited effect of systematic phonics instruction on word reading (e.g. Porter, 2020; Woore et al., 2018) and it is not always clear whether improvement in literacy skills is due to increased phonological decoding skills specifically. This lack of supporting evidence is, in part, attributable to a paucity of reliable phonological decoding assessment tools that can be used with a large enough sample to ensure sufficient power in intervention studies. Relatedly, a recent scoping review (Liang & Fryer, 2024) of phonics and phonemic awareness instruction in East Asian EFL contexts concluded that these had a positive impact on code-related skills (e.g. decoding), oral language and reading comprehension skills. Yet little attention was paid to how phonological decoding or indeed other outcomes were assessed in the studies, except to determine whether a standardised or non-standardised test was used. Examination of YL studies included in the review also revealed little real detail of instruments for assessing decoding or scoring procedures, alongside relatively small sample sizes (Chu & Chen, 2014; Ng, 2006).

In short, whether the arguments made in support of phonological decoding enrichment in L1 contexts extend to FL contexts is unclear. In order to analyse the link between FL phonological decoding skills and FL reading outcomes among YLs, a necessary first step is to devise a reliable means of evaluating YLs' FL phonological decoding. This assessment strategy must be sensitive not only to the subtle changes that such learners are likely to undergo, but also to the important differences between L1 decoding and FL decoding.

2.3. Considerations for the assessment of FL phonological decoding vs. L1 decoding

While there are many processes that both first and second language (L2) acquisition share, there are also crucial differences. For example, for sequential L2 learners, by the time the second language is encountered the bilingual is older, cognitively more advanced and socially more aware (Cook, 2010). One of the most important aspects of FL phonological decoding that differentiates it from L1 decoding is the pre-existing knowledge of another sound system and its writing system(s). Unlike assessment of phonological decoding in a first language, then, FL phonological assessment must take into account that FL phonological decoding involves suppressing, adapting or transferring SSCs already established between symbols and sounds in the L1 (Bassetti, 2024; Hayes-Harb & Barrios, 2021).

Assessment of FL phonological decoding should also differ substantially from assessment of reading in a first language because in the latter, children will likely know or will have heard test items before that they are attempting to decipher. That is, in L1 contexts, the

[2] We acknowledge, however, that for some SSCs, the app's scores may be more highly representative of a pronunciation difficulty than a decoding issue *per se*.

process is more analogous to word recognition: a number of potential candidates from stored lexical representations are activated while reading from which the target is recognised (see Hendrickson et al., 2022 for an overview of written and auditory word recognition). In contrast, FL words are unlikely to be as strongly encoded among YLs and will often not have been encountered before, especially if the word has a low word frequency (Dudley & Marsden, 2024). To assess the acquisition of SSCs in L1 reading contexts then, non-word test items are often thought necessary to ensure that phonological decoding is sublexical and that whole-word recognition strategies are not employed (Gibson & England, 2016). For instance, the Phonics Screening Check (Standards & Testing Agency, 2024) is a legal requirement in Year 1 of primary school (5–6 year olds) in England and involves the reading and scoring of non-words. However, the motivation for using non-words to tap sublexical decoding may have less relevance for assessing FL phonological decoding among YLs, given that most words will either be new or less familiar. In such instances, reading is slower, more cognitively exhausting and long tasks run the risk of obtaining inaccurate data from performance fatigue. As such, assessment methods for FLs must be short enough to hold the attention of the user, a consideration especially important for YLs (McKay, 2005), yet long enough to obtain reliable data.

The assessment of phonological decoding in an FL is also crucially different from L1 phonological decoding if the transparency of the orthography in the FL differs from the L1. For example, languages such as English and French have "deep" orthographies meaning that symbols may be pronounced in a variety of ways. Indeed, in some dialects of English there are at least seven ways in which *ough* can be pronounced in words such as *rough, through, cough, though, thought, plough* and *borough*. In other languages, such as Spanish, orthography is much more "shallow": often a single symbol will correspond to one and only one sound. Generally, decoding is arguably easier in languages with transparent orthographies (Cable et al., 2010). However, according to the psycholinguistic grain size theory (Ziegler & Goswami, 2005), children whose L1 has a deep orthography, such as English, may become reliant on decoding longer sequences of graphemes or spelling bodies (Egan et al., 2019; Lallier & Carreiras, 2018). As such, they may possess strategies that transfer more easily to deep orthographies (e.g. French) but may be unaccustomed to decoding smaller units used in shallow orthographies (e.g. Spanish). Therefore, it is useful to validate multi-language FL phonological assessment tools using language pairs of differing levels of consistency in terms of orthographic depth.

2.4. Challenges in assessing FL phonological decoding

The assessment of FL phonological decoding, hence, poses many challenges, especially in YL classrooms. One of the main issues is the lack of a well-validated phonological decoding assessment task. Orthographic processing tasks exist (orthographic/homophone/wordlikeness choice tasks, see e.g. Apel (2011) for an overview) but these often do not measure knowledge of individual SSCs. Instead, read-aloud tasks are a common method in research in this domain (see e.g. Hamada & Koda, 2008; Porter, 2020). In these tasks, as in L1 tasks (see Section 2.3), words or non-words are presented and the participant is instructed to read the item aloud (decode), and auditory coding by a human assessor will evaluate whether items are accurate or inaccurate. Woore et al. (2018) suggest tasks of this type are particularly sensitive compared to alternatives such as sound-alike tasks in which participants see multiple written items and select those that sound alike when read aloud. Potentially this is because sound-alike tasks are more cognitively challenging, especially for YLs for whom the level of task complexity is a particularly important consideration (McKay, 2005). For instance, in the course of just one question on a sound-alike task, multiple items (often three) must be a) decoded, b) stored in short term memory, and c) compared with each other, before a subsection of items are ultimately selected. Yet the assessment of read-aloud tasks is not always feasible with large sample sizes (Erler & Macaro, 2011; Woore, 2021; Woore et al., 2018). Indeed, assessing FL phonologically decoding is extremely labour intensive to administer and score, a fact that also holds true for pronunciation assessment (De Jong, 2023). Manual phonology assessment of large numbers of learners is impracticable, not only for researchers but also for teachers, leading scholars to call for "creative solutions" (Woore, 2021, p. 240) to deliver a suitably sensitive test without overloading those needing to implement it.

Workload is not the only difficulty for manual assessment of phonological decoding. Teachers of YLs may have lower linguistic proficiency levels than could be needed for such a task (Collen & Duff, 2024; Unsworth et al., 2015). If teachers, or indeed anyone undertaking a manual assessment, lack a stable phonology in the target FL, their perception of accuracy will be heavily guided by their L1 sound system and their own non target-like L2 sounds, as has been shown frequently in L2 speech perception research (Best & Tyler, 2007; van Leussen & Escudero, 2015). This is likely to impact their assessment accuracy in turn (Myford, 2012). Furthermore, when scoring in real time, raters must be able to quickly process and accurately recall L2 speech which will largely depend on a number of factors including language proficiency, cognitive load and working memory (Han, 2016). Automated assessment can help to reduce many of these concerns as well as facilitating large-scale assessment, hitherto deemed impossible.

2.5. AI and automated speech assessment

Unlike assessment of writing (see e.g. Geçkin et al., 2023; Pfau et al., 2023), few studies to date have used AI to assess L2 speaking or for speaking instruction. For example, in a recent meta-analysis of AI-assisted language learning, none of the 35 studies reported analyse spoken data (Lee & Lee, 2024). Furthermore, it has been argued that "for classroom assessment […] automatically measuring speech is currently not feasible" (de Jong, 2023), potentially due to a lack of suitable tools to do so. The present study explores the extent to which this remains true in light of growing evidence of a positive role for computer-assisted language assessment, along with recent advances of AI.

One concern for AI-backed speech assessment, such as tools that use automatic speech recognition, is that the scoring system may not perform consistently across different recordings due to external factors such as background noise (Litman et al., 2018). Indeed, sensitivity to the recording environment is a well-known issue in the domain of automatic speech recognition, and is a variable that must, therefore, be reduced or controlled for (Benzeghiba et al., 2007). Hence, for an AI tool to be successful in automating assessment, automated scores should arguably not only correlate with human coding (Litman et al., 2018), but also be consistent across measures of environment noise like signal-to-noise ratio (Litman et al., 2018), or, at least, affected to a similar extent as is human coding. The AI's scoring must also be sensitive to the methodological variables discussed in Section 2.3: phonological decoding of languages with greater orthographic depth are likely to yield lower values (Cable et al., 2010); real words are expected to have higher scores than non-words if the real words are already known to learners (Gibson & England, 2016); and language learners who are more proficient in the target language are likely to achieve higher scores (Hamada & Koda, 2008; Jeon & Yamashita, 2014). Nevertheless, these variables are expected to affect both human auditory coding and the AI automated scores to a similar extent. As such, associations between human coding and AI generated scores should be relatively consistent across variables such as the orthographic depth of the language, word type, and learners' proficiencies in the language.

In short, whether modern AI architectures such as transformer models that employ DNNs trained on big data can be used for the purposes of assessing FL phonological decoding among young FL learners is an important line of investigation, especially given the scalability of many AI tools and their capacity to circumvent human subjectivity. Nevertheless, studies focusing on this line of investigation should also evaluate whether the relationship between human and automated coding is consistent across variables that have a known influence on FL phonological decoding. If a successful method can be established, there are important follow-on implications for studies seeking to make judgements on the effectiveness of phonological instruction for reading and other aspects of L2 learning.

2.6. The present study

This research analyses the feasibility of large-scale data collection and language assessment in classroom contexts using an AI-integrated app for tablets, mobile devices and ChromeBooks. We report on an app-assessed FL phonological decoding (read-aloud) task in UK primary schools that was administered to learners in five schools. In an initial validation of the app's scoring system, 14 symbol-sound correspondences (7 in two languages) were tested by comparing the scores generated by the app and human auditory coding of SSC accuracy. Given that L2 pronunciation research has placed less importance on assessing "foreign accentedness" or "nativelikeness" in recent decades (see e.g. Levis, 2018; Saito, 2021), the present research also establishes the extent to which word-level scores generated by the app corresponded to human coding for "comprehensibility", namely, how difficult it is to understand the word as a whole. Given that the app's word level scores are computed by comparing user utterances to native speaker recordings (Section 3.4), it is worth evaluating to what extent these scores are nonetheless associated with human ratings of comprehensibility, a somewhat distinct, but related, construct (Levis, 2018).

In both analyses (namely, sound and word levels) we further explored the extent to which correlations between app and human scores were modulated by the linguistic proficiency of the learner, as indicated by reading comprehension and vocabulary in the FL, the language itself (French or Spanish) and the quality of the recording (Harmonics-to-noise ratio: HNR). In addition, the sound-level analysis tested the effect of word type (real word vs. non-word) but not the word-level comprehensibility analysis because the latter only included real words (see Section 3.5). Our research questions were therefore as follows:

SSC-level Analysis (real word and non-words)

RQ1) a) to what extent do the app's scores predict sound-level human auditory SSC accuracy coding? b) is this relationship modulated by individual variation in linguistic proficiency (two proxy measures: total reading comprehension score and average receptive vocabulary score), or other methodological factors such as Word Type (real word vs. non-word), Language (French vs. Spanish) or HNR (a proxy for recording quality)?

Word-level Analysis (real words only)

RQ2) a) to what extent do the app's scores predict word-level human auditory coding for comprehensibility? b) is this relationship modulated by individual variation in linguistic proficiency (two proxy measures: total reading comprehension score and average receptive vocabulary score), or other methodological factors such as Language or HNR?

3. Methodology

3.1. Participants

The phonological decoding of 254 FL YLs was analysed in this study: 128 of these learners were L1 English speakers learning French at a UK primary school and 126 were L1 English speakers learning Spanish (both groups mixed genders, ages 10–11, third year of L2 education). Participating students formed part of a broader literacy intervention research project involving 10–11 year-old primary students in the UK, but the present study limits its focus to the relationship between the app scores and the manual auditory coding. Both total reading comprehension scores and average vocabulary scores across the sample were used as proxy variables for linguistic proficiency (see Sections 3.4 and 3.6). The vocabulary results revealed an overall mean of 23/40 (range: 7–38, SD: 5.7), indicating these were low to intermediate learners comparable to those in other YL studies (e.g. Courtney et al., 2017).

Table 1
SSCs investigated by this study.

Target SSC Variable	Language	Example in FL	Possible English SSC transfer	Example in L1
<ch> – [ʃ]	French	**ch**ien (dog)	<ch> – [t͡ʃ]	**ch**in
<j> – [ʒ]		**j**e (I)	<j> – [d͡ʒ]	**j**am
<on> – [ɔ̃]		t**on** (your)	<on> – [ɒn]	b**on**d
<ou> – [u]		l**ou**p (wolf)	<ou> – [ʉː] / [aʊ]	gr**ou**p/h**ou**se
<u> – [y]		t**u** (you)	<u> – [ʉː] / [jʉː] / [ʌ]	d**u**de/red**u**ce/b**u**s
<qu> – [k]		**qu**and (when)	<qu> – [k^{h}w̥]	**qu**een
<r> – [ʁ]		**r**at (rat)	<r> – [ɹ]	**r**ight
<j> – [x]	Spanish	**j**unio (June)	<j> – [d͡ʒ]	**j**am
<ñ> – [ɲ]		a**ñ**o (year)	<n> – [n]	a**n**imal
<qu> – [k]		**qu**itar (to take off)	<qu> – [k^{h}w̥]	**qu**een
<r> – [r]		**r**ápido (fast)	<r> – [ɹ]	**r**ight
<u> – [u]		m**u**cho (many)	<u> – [ʉː] / [jʉː] / [ʌ]	d**u**de/red**u**ce/b**u**s
<v> – [b]		**v**aca (cow)	<v> – [v]	**v**an
<z> – [θ]		**z**apato (shoe)	<z> – [z]	**z**ip

3.2. *App functionality*

The AI-based app designed for this project functions for both French and Spanish and on both *iOS* and *Android* devices. It presents visual, written stimuli to the user to elicit spoken responses which are then recorded by the app and stored to the research project's server. Using a compressed neural network, the app generates a phonetic transcription of any utterance within milliseconds. Using this transcription, the app calculates a number of scores (Section 3.4) and can provide real-time scoring and narrative feedback to the user, although this functionality was deactivated for the data used in the present study. The task was broken down into short subsections and at the end of each, the children were able to unlock customisations for their character avatars, including a variety of facial features, hair styles, and accessories. This gamification aspect ensured that enjoyment levels remained high while participating (Courtney & Graham, 2019; Lampropoulos et al., 2022). It also allowed these YLs to take a suitable number of breaks: after producing one subsection of items, a new set of avatar customisations became unlocked (see McKay, 2005 on the importance of age-appropriate assessment). This process is demonstrated in Appendix A.

3.3. *Test instrument*

Participants undertook a read-aloud task (RAT) within the app which consisted of 14 items (Appendix B). Seven symbol-sound correspondences (SSCs) were investigated in each language (i.e. two items each). For each SSC, one item was a real word while one was a non-word. Real words were included to facilitate obtaining ratings of comprehensibility in the auditory coding; these are less likely to be reliably obtained if a rater does not already know the words that they are required to recognise. At the same time, including non-words ensured that decoding was sublexical because it would not be the case that participants had already encountered this item in the FL. Furthermore, the real words that were chosen were low frequency (below the top 7000 most frequent tokens in the French Lexique corpus (New et al., 2004) or Spanish SUBTLEX corpus (Cuetos et al., 2011)) to reduce the likelihood that they were already known by participants. Other criteria for the real words were that they were made up of a maximum of two syllables to reduce complexity, that they contained only one of the SSCs analysed in the study, and orthographic exact cognates (or false-cognates) in relation to the L1 (English) were excluded. Almost-exact cognates were also avoided as far as possible to reduce their influence on results.

The SSCs themselves were chosen because of their relatively high rates of word-internal consistency in the FL as determined using a combination of pronunciation dictionaries (McAuliffe & Sonderegger, 2022) and grapheme-phoneme corpora where available (e.g. Infra-Lexique, Gimenes et al., 2020). We argue that learning these consistent SSCs is a suitable starting point for YLs given that learning less predictable symbols (that correspond to many different possible sounds) will inevitably be a more complex learning task. Consistency was measured on a position-sensitive allophone basis. For example, Spanish <v> – [b], French <r> – [ʁ] and Spanish <r> – [r] were only assessed in word-initial position as other contexts may yield production of different allophones in the target languages. Although not all SSCs are 100 % consistent even in the restricted positions, we attempt to focus on the dominant sound for each symbol in the context provided. Table 1 shows the target 14 SSCs, along with the links between letters and sounds that are likely to transfer from L1 English. These SSCs fall into two groups: SSCs with target FL sounds that overlap phonetically with L1 English (e.g. <ch> – [ʃ] in French and <z> – [θ] in Spanish: [ʃ] and [θ] are present in English, e.g. *ship* and *thin*); and those that have less phonetic overlap (e.g. <r> – [ʁ] and <r> – [r]). Furthermore, for this latter group of SSCs involving articulatorily new sounds, some symbols encourage transfer of the nearest L1 sound (e.g. <on> for [ɔ̃] is a SSC in L1 English for the most similar L1 sound sequence, oral [ɒn]), while others do not (e.g. for Spanish <j> – [x], [x] is a new sound that is often perceptually assimilated to English [h], but the symbol <j> does not encourage the use of English [h] because <j> – [h] is not a SSC in English. In a separate study, we analyse whether these different types of SSC hinder or facilitate development of phonological decoding over time but for the purposes of the present study, we restrict our focus to the validation of the AI app's scoring system rather than evaluating SSC difficulty.

The task was administered in the language taught at the school (French or Spanish, not both). The app presented test items individually in a pseudo-randomised order with a run controller preventing the same SSC being presented twice in a row and this order was fixed across participants. In the RAT, the item's orthographic form was presented for an unlimited amount of time with no audio and participants were asked to read the visual item aloud.

3.4. Procedures and app scoring

In order for the app to compute scores for individual recordings, a number of steps were necessary. First, the expected transcriptions of the target forms in the RAT were extracted or adapted from the *Montreal Forced Aligner* (McAuliffe et al., 2017) dictionaries in both French and Spanish (McAuliffe & Sonderegger, 2022). Next, a production of the target items was elicited from four L1 speakers of the FLs (two French, two Spanish) with genders balanced within language. These recordings were run through the neural network integrated into the app in order to produce a set of International Phonetic Alphabet (IPA) transcriptions for each item. These transcriptions were stored alongside the target forms extracted from the Montreal Forced Aligner pronunciation dictionaries. These exemplars, or "ground truth" transcriptions, would then be used for the scoring mechanism (see below).

Teachers in the five schools provided tablets in FL classes to their students with the phonics app already installed. Participants would log in to the app via a username and password distributed in advance and commence the decoding task. The app processed all speech recordings through local AI inference: the DNN model generated an IPA transcription and pronunciation scoring was compressed and optimised to be stored on the user's device and performed inference in real-time. All speech recordings, IPA transcriptions and scoring data were then transmitted to a secure remote server that could later be reviewed by the research team. A multilanguage model was used which meant that the resultant transcriptions had the potential to include all IPA symbols rather than just those of the target language. The similarity between the phonetic transcription of the item and the ground truth transcriptions was immediately calculated both in terms of the word as a whole (in-built app word-level sound score) and for specific sounds within the item (in-built app SSC score). Because there was often more than one ground truth transcription that a user utterance was compared against, only the highest of the similarity scores was used. These scores were gradient (between 0.0 and 1.0) and were not consulted in any detail before the auditory coding stage and neither these scores nor any feedback were provided to the participants in this testing phase. A final word-level "overall" score was also generated by the app which was, once again, a value between 0.0 and 1.0 corresponding to how similar the user's utterance was to the ground truth transcription along a number of dimensions including sound quality, pitch, amplitude, segment timing, stress and overall length. This differs from the word-level sound score which only took into account accuracy of the sounds within the word.

In addition to the app's integrated SSC score, a secondary accuracy metric was calculated by the research team which determined whether the target sound appeared in the correct position in the app's neural network IPA transcription of the participant's utterance. Unlike the app's gradient score described above, this post-hoc app score was binary: if the appropriate sound was observed in approximately the correct position (that is, the correct index ± 1 place), the item was scored as "1" or "0" otherwise.

After excluding recordings where no speech was produced, across the five schools a total of 3017 tokens were coded in the SSC analysis (1509 Spanish, 1508 French) and a total of 1585 tokens were considered in the comprehensibility analysis (820 for Spanish and 765 for French).

In addition to the decoding tasks, participants also undertook a bespoke reading comprehension task developed for the project that comprised both translation and multiple-choice questions (Cronbach's alpha = 0.94) and a vocabulary knowledge test. Vocabulary was measured using the 40-item test developed and validated by Morea et al. (2024) on primary school learners in England. This is a test of meaning recognition in which learners hear an item in French or Spanish and select its meaning from four options given in English. Rasch analysis showed the person reliability coefficients of the French and Spanish version of the test to lie within an acceptable range between 0.72 for the latter and 0.80 for the former (Morea et al., 2024) indicating that the test has a good ability to differentiate between individuals based on their vocabulary levels. Total scores for vocabulary and reading comprehension were entered, separately, in statistical models as proxy measures for linguistic proficiency (see Section 3.6).

3.5. Auditory coding

The recordings were first downloaded from the app's server. Each recording was then coded manually using a script written by the first author in *Praat* (Boersma & Weenink, 2024). This script also generated the mean Harmonics to Noise Ratio (HNR) across the length of each recording which, as mentioned in Section 2.6, was used as a proxy for recording quality. The real lexical items were coded for comprehensibility using a 4-point Likert scale (1: this item is very difficult to understand, 2: this item is quite difficult to understand, 3: this item is quite easy to understand, 4: this item is very easy to understand) and both the real and non-words were assigned an accuracy score for the target grapheme-phoneme correspondences (1: right sound or very little influence from L1, 0: wrong sound or clear influence from L1). Influence from the L1 was only analysed so that coders would not avoid giving an accurate score if the pronunciation was not nativelike as this was deemed an unrealistic target (Levis, 2018). This is a key difference from the app's scores – which are based on similarity to native speaker recordings – thus warranting an investigation into the association between the two methods. The primary coder (the first author) is an L1 English speaker and a trained phonetician with a background in L2 speech and advanced knowledge of French and Spanish. The coding was repeated for 10 % of the data by another of the authors and this sample was stratified across SSC variable and language to ensure a variety of SSCs and a proportionate number of tokens from both languages were reanalysed. These interrater reliability checks indicated sufficient overlap in assigned codes: there was 81 % agreement between raters for SSC codes and Cohen's Kappa was 0.58 indicating "moderate" agreement, and in the comprehensibility analysis, a Spearman's rank correlation revealed a significant relationship between coders: $r = 0.64$, $p < .001$. Weighted (quadratic) Cohen's Kappa was 0.33 suggesting a "fair" level of agreement) (Landis & Koch, 1977).

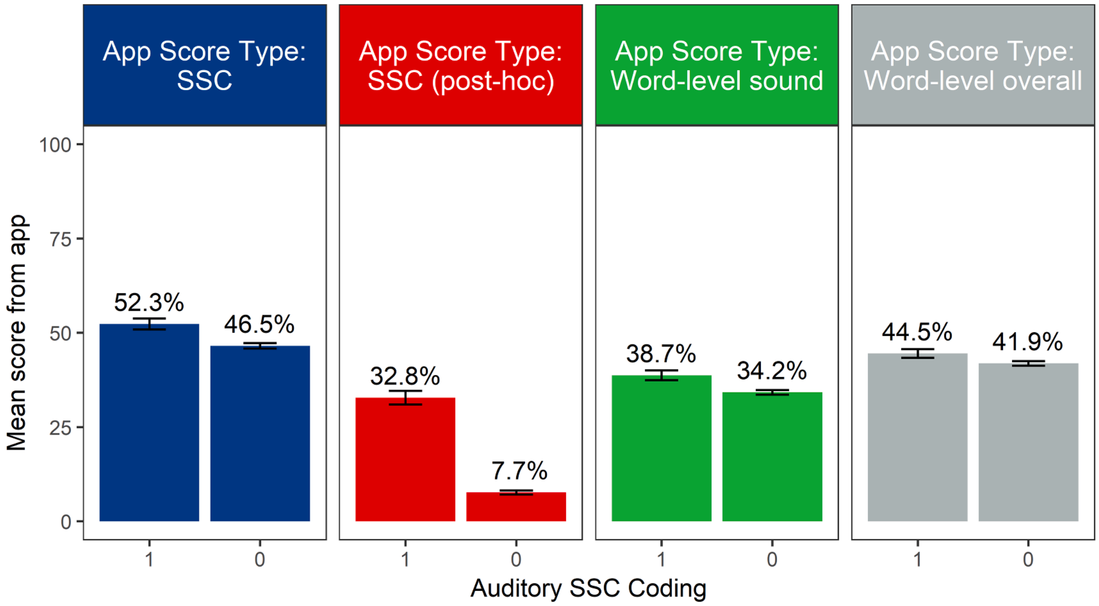

Fig. 1. The means of the 4 app scores (y-axis) coded as accurate (1) in the auditory SSC coding and inaccurate (0) (x-axis).
Note: In blue: in-built sound-level SSC score; in red: SSC score calculated post-hoc from neural network transcription; in green: sound score for the entire word; in grey: overall word score (accounts for sound, pitch, timing, length, stress and amplitude).

3.6. Statistical analysis

The first analysis, focusing on the 14 individual SSCs, involved a series of mixed effects binomial logistic regression models with the auditory code (1 vs. 0) as the dependent variable. These models were fitted using the *lmerTest* (Kuznetsova et al., 2020) and *lme4* (Bates et al., 2015) packages in *R* (R Core Team, 2024). In separate models, the in-built app SSC score, the post-hoc SSC score, the app word-level sound score and the app overall word score (see Section 3.4) were tested as fixed effects. Variables such as vocabulary and reading comprehension scores (scaled), Language (effects coded), SSC Variable (deviation coded), Word Type (effects coded) and HNR (scaled) were also tested as main effect and as interactions with the app scores listed above. This was to establish whether the relationship between the app score and the auditory coding was modulated by these variables.

The second analysis, focusing on word-level scores, involved a series of mixed effects ordinal logistic regression models with the ordered auditory comprehensibility codes (1–4) as the dependent variable. These models were fitted with the *ordinal* package (Christensen, 2023) once more in *R* (R Core Team, 2024) with flexible thresholds and the same fixed effects as above were tested once more. The exception was Word Type (as the comprehensibility analysis focused on real words only).

In all models (logistic and ordinal) random intercepts were included for Participant and Item and maximal random slopes for all main effects and interactions (Barr et al., 2013). This random structure was only reduced in instances with convergence issues. School was included as a further random intercept for the ordinal models after a likelihood ratio test suggested it significantly improved model fit. This was not found to be the case for the binomial logistic regression models so Participant and Item were the only random intercepts in these models. The significance of fixed effects was determined using likelihood ratio tests through the *drop1* function (Kuznetsova et al., 2020). Where pairwise comparisons of estimated marginal means are reported, these are post-hoc estimates calculated using the *mvt* adjustment method in the package *emmeans* (Lenth et al., 2024). Emmeans output can be found in the Appendices and the data used in this study are publicly available: https://osf.io/2kxdv.

4. Results

4.1. SSC analysis

to what extent do the app’s scores predict sound-level human auditory SSC accuracy coding?

We begin by presenting descriptive statistics (Fig. 1).

Fig. 1, focusing on individual symbol-sound correspondences, demonstrates that the means of all four scores generated by the app were numerically higher for tokens coded as accurate in the auditory coding than those coded as inaccurate. This is particularly the case for the SSC score calculated post-hoc (red), and to some extent the in-built app SSC score (blue). There was a much smaller difference in scores between accurate and inaccurate tokens for the app’s word level scores, however. This was expected because the word-level scores take into consideration many more aspects of the pronunciation than the specific singular SSC embedded within the item.

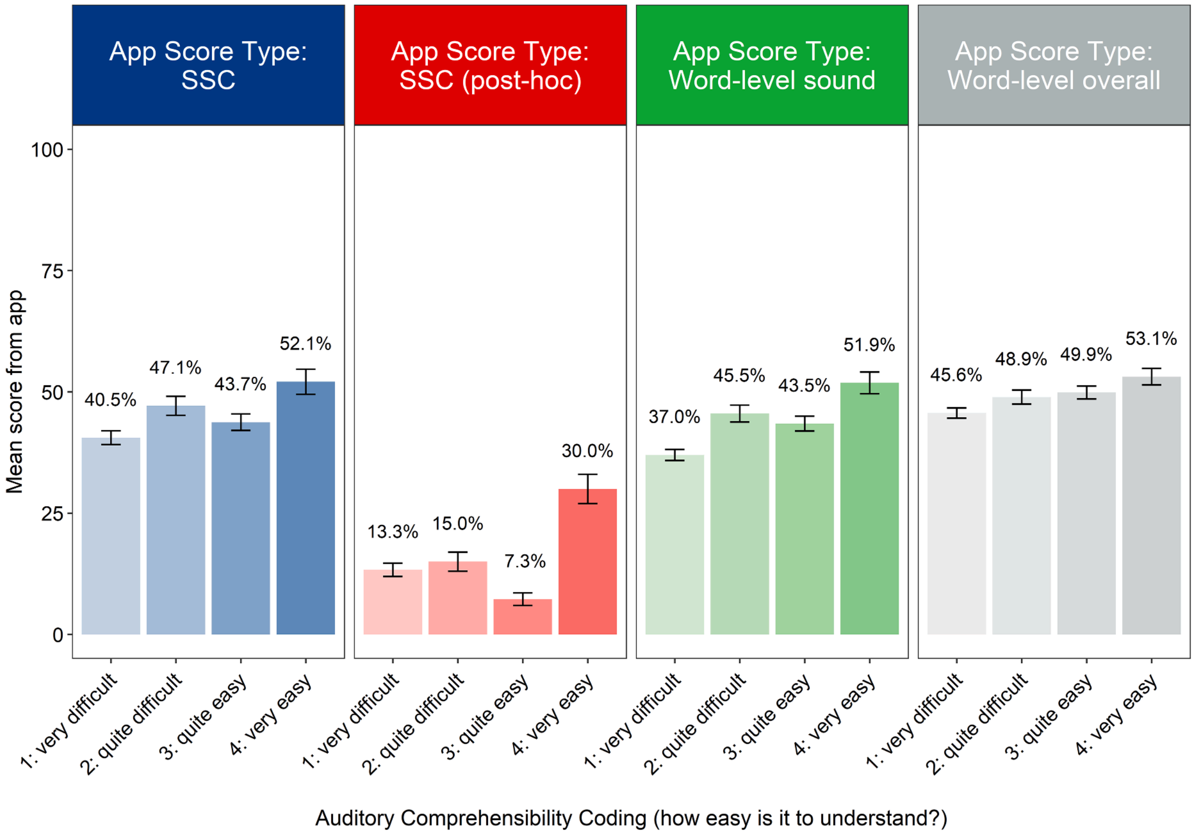

Fig. 2. Means of the four app scores for words coded as 1: very difficult to understand, 2: quite difficult, 3: quite easy and 4: very easy. Note: In blue: in-built sound-level SSC app score; in red: SSC app score calculated post-hoc from neural network transcription; in green: app sound score for the entire word; in grey: app overall word score (accounts for sound, pitch, timing, length, stress and amplitude).

To investigate further, individual binomial logistic regression mixed effects models were fitted with the auditory coding variable (0, 1) as the dependent variable and the four app scores fitted as fixed effects (as detailed in Section 3.6). Results revealed that both the in-built App SSC score $\chi^2(1, N = 2999) = 7.19, p = .007$ and the App SSC score calculated post-hoc $\chi^2(1, N = 3017) = 12.57, p < .001$ significantly predicted the human auditory coding. That is, the likelihood of an accurate auditory code increased in line with both the app's in-built SSC score (β = 0.20, SE = 0.07, $z = 2.77$ $p = .006$), and the app's post-hoc SSC score (β = 1.84, SE = 0.37, $z = 4.95$, $p <$ .001). No significant effect of the app's word-level scores was observed on the auditory SSC coding, however, as expected.

4.2. SSC analysis

is the relationship between app scores and auditory coding modulated by individual variation in linguistic proficiency or methodological factors?

To establish whether these relationships were modulated by individual variation in linguistic competence, other methodological factors, participants' vocabulary score, reading comprehension score, Language, SSC Variable, HNR and Word Type were tested as interactions with all four app scores. A significant interaction was observed between the app's post-hoc SSC score and the SSC Variable $\chi^2(11, N = 2588) = 34.04, p < .001,$[3] suggesting the SSC variable modulated the relationship between the app scores and the auditory coding. Given that no other significant interactions were observed, variables such as linguistic proficiency (reading comprehension and vocabulary), word type, language and HNR did not appear to modulate the relationship between the app scores and the auditory coding. The sole significant interaction showed that the relationship between the app's post-hoc SSC score and the auditory coding was not consistent across the 14 SSCs; pairwise comparisons of estimated marginal means (Appendix C) further revealed that the post-hoc SSC app score mirrored the auditory SSC coding most closely for <j> – [x], <z> – [θ], <r> – [r] and <u> – [u] in the Spanish data and <ch> – [ʃ], <j> – [ʒ] and <ou> – [u], in the French data.

[3] This analysis excluded the SSC variables <r> – [ʁ] and <on> – [ɔ̃] in French as it transpired that none of the tokens containing these SSCs were accurate in the post-hoc App SSC Score (the phonetic transcription did not contain the target sound in approximately the correct position), thus there was not sufficient variation to fit an interaction.

4.3. Comprehensibility analysis

to what extent do the app's scores predict word-level human auditory coding for comprehensibility?

We again began by producing descriptive statistics (Fig. 2).

Fig. 2 (two right-hand panels) suggests that the app's word-level scores increased incrementally with the Likert scale for comprehensibility in the auditory analysis. That is, as the ease in understanding the word increased in the auditory coding, so did the app's word-level scores. The sound-level SSC app scores also appeared to correspond to the word-level auditory coding for comprehensibility, though the incremental increase was slightly less clear than the word-level app scores.

Next, the data were analysed statistically using mixed effects ordinal logistic regression with the auditory coding fitted as a 4-level ordered dependent variable and the app scores fitted as fixed effects (as detailed in Section 3.6). Results revealed that the auditory coding of comprehensibility for the real words was predicted by both the app's word-level sound score $\chi^2(1, N = 1585) = 8.69, p = .003$, and the app's word-level overall score $\chi^2(1, N = 1585) = 4.64, p = .031$. Inspecting the estimates revealed that as both word-level app scores increased, so did the comprehensibility of the word according to the auditory coding (app word-level sound score, β = 0.27, SE = 0.08, $z = 3.53, p < .001$; app word-level overall score, β = 0.18, SE = 0.08, $z = 2.37, p = .018$). The word-level auditory coding for comprehensibility was also predicted by the app's post-hoc SSC score: $\chi^2(1, N = 1585) = 5.65, p = .017$ but not the app's in-built SSC score: $\chi^2(1, N = 1564) = 1.67, p = .196$. That is, as the app's post-hoc SSC scores increased, on average, so did the comprehensibility of the word according to the auditory coding (β = 1.33, SE = 0.48, $z = 2.77$, $p = .006$).

4.4. Comprehensibility analysis

is this relationship modulated by individual variation in linguistic proficiency or methodological factors?

When participants' vocabulary score, reading comprehension score, Language, SSC Variable, and HNR were tested as interactions with the four app scores, the only significant effects were between the two SSC app scores and the SSC Variable (for the in-built app SSC score: $\chi^2(13, N = 1564) = 22.79, p = .044$; and for the post-hoc app SSC score: $\chi^2(11, N = 1363) = 34.60, p < .001$). This demonstrates that the relationship between the app's in-built SSC score and the human-coded comprehensibility scores differed across words containing each SSC, and that this was also the case for the app's post-hoc SSC score. Indeed, pairwise comparisons of estimated marginal trends (Appendix D) revealed that the relationship between the in-built app's SSC scores and the word-level human coding for comprehensibility was strongest for the following SSCs: <ch> – [ʃ]; <j> – [ʒ] and <r> – [r]. Similarly, pairwise comparisons of estimated marginal trends (Appendix E) revealed that the relationship between the app's post-hoc SSC scores and the word-level human coding for comprehensibility was strongest for <ch> – [ʃ], <j> – [ʒ] and <r> – [r], but also <j> – [x], <z> – [θ], and the French <qu> – [k]. Given that none of the other variables interacted with the app scores, the relationship between the app scores and the auditory coding appeared somewhat independent of linguistic proficiency, language (orthographic depth), and differences in recording quality.

5. Discussion

This study proposes an AI-based app as an assessment method with great potential for providing insight into YLs' FL phonological decoding and, hence, a methodological innovation for research. Our focus, here, was to provide an initial validation of the app's generated scores by examining the correspondence with the auditory coding performed by an expert rater.

Our first research question asked whether any of the four app scores were significant predictors of the sound-level human auditory coding for SSC accuracy. Results revealed the two sound-level app scores (one in-built ranging from 0.0 to 1.0, and one binary variable calculated post-hoc by the research team referring to the presence or absence of the target IPA symbol) were significantly associated with the SSC auditory coding while the two word-level app scores were not, as expected. This suggests that when analysing learners' capacity to decode individual SSCs into intelligible sounds, the app's sound level scores are, at least to some extent, effective. This is particularly noteworthy given the in-built app score is based on similarity to native speaker exemplars while the auditory coding places less importance on nativelikeness. Given that the app's word level scoring was designed to be a general score across all phones and phonological feature dimensions spread over all sounds and features in a word, it was not unexpected that there would be little to no correlation between word-level app scores and the human auditory coding that focused on an individual sound. Further analyses revealed that the relationship between the post-hoc SSC app score was stronger for some SSCs than others, but that the relationship between auditory coding and the app SSC scores did not vary substantially between different linguistic proficiency levels (at least in terms of vocabulary knowledge and reading comprehension), different word types (real words vs. non-words), languages (French vs. Spanish) or HNR (a proxy for recording quality). This suggests the app's SSC scoring performs relatively consistently across different learners and contexts.

Our second research question asked whether any of the four app scores were significant predictors of the word-level human auditory coding for comprehensibility. Results revealed that as both the word level scores increased, so did the comprehensibility ratings. Furthermore, the app's post-hoc SSC score was found to predict the human coder's ratings for word comprehensibility. Therefore, if the user's aim is to decode text into a comprehensible utterance or, indeed, for an assessor to understand how well a user performs in this respect, any of these three scores are likely to prove reliable. This is particularly interesting given the somewhat distinct approaches taken by the human coder (under which native similarity was not of primary importance) and the app's scoring system (which focused on similarity to native exemplar recordings). Again, further analyses revealed that the relationship between the comprehensibility scores and both a) the post-hoc SSC app score and b) the in-built SSC app score was modulated by the SSC variable in

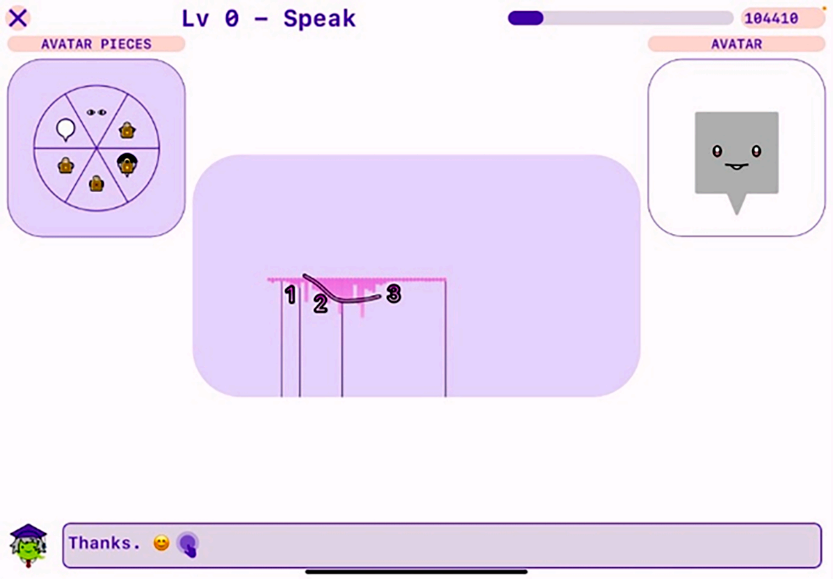

Fig. 3. *Screenshot from app for reading aloud "quime".*

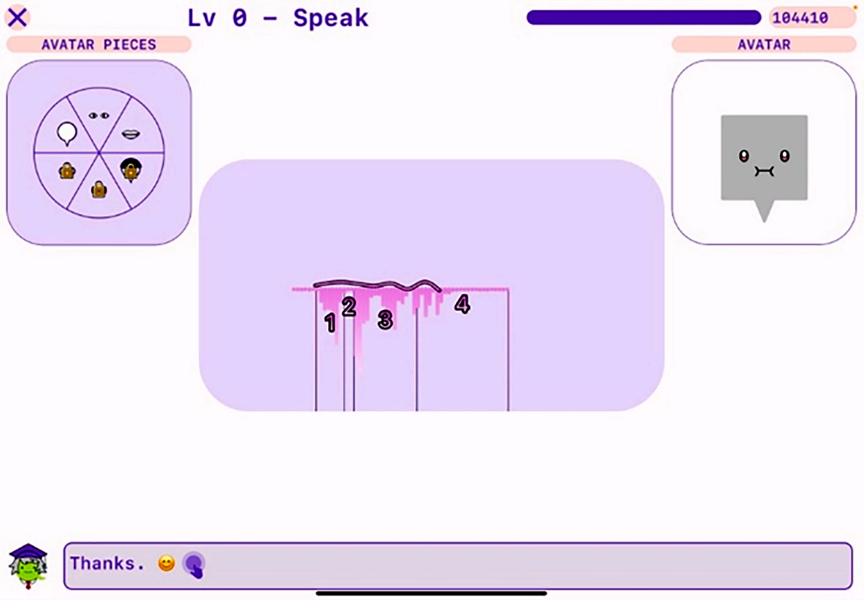

Fig. 4. *Screenshot from app for reading aloud "abondent".*

question. Nevertheless, the relationship between the app scores and the comprehensibility ratings appeared to be consistent across reading comprehension scores, vocabulary scores, language and HNR.

To our knowledge, this is the first study to demonstrate the viability of using AI to assess FL phonological decoding within this population. This marks a clear advancement of the field and has important implications not only for language learning and assessment research but also classroom practice and policy. For instance, as these types of phonological decoding assessment tools advance, future research will be able to address questions at the forefront of YL FL education such as the extent to which phonological decoding development in an FL has a positive influence on FL literacy outcomes, FL pronunciation and fluency and FL processing. The app used in the present research is especially appropriate for this target audience (YLs); with only 14 items and gamification aspects embedded, including avatar development and customisation, tasks are less cognitively taxing, more engaging (Lampropoulos et al., 2022), and more age-appropriate for language research on YLs (Courtney & Graham, 2019; McKay, 2005). By facilitating research into YL FL phonological decoding, we hope to prevent existing *first* language reading research and policy being generalised to FL contexts without extensive prior investigation concluding it is appropriate to do so. This is especially important given that studies of FL reading among YLs do not appear to paint the same picture as first language contexts with respect to the link between phonological decoding and learning to read (Porter, 2020; Woore et al., 2018).

The depth and scale of the data sets that can be obtained from this technology offer robust evidence for answering such questions. The current study joins a wealth of recent literature advocating for the use of AI in different aspects of language assessment (Geçkin et al., 2023; Pfau et al., 2023), thus ensuring studies are not limited in scope to one small sample or specific context. Further, the app responds to recent calls from the field (Woore, 2021) to use sensitive measures of FL decoding among YLs such as read-aloud tasks (Woore et al., 2018) while simultaneously avoiding the confound of teacher L2 linguistic proficiency in assessing students (Collen & Duff, 2024; Unsworth et al., 2015), the cognitive factors influencing between-assessor consistency (Han, 2016), and the unfeasible assessment of large quantities of speech (de Jong, 2023; Woore, 2021; Woore et al., 2018).

These advantages extend to classroom practice. For example, the app can avoid additional workload being placed on practitioners by overcoming the need to assess students on a one-to-one basis. Excessive workload exacerbates broader issues currently affecting education such as teacher stress and mental health wellbeing (Agyapong et al. 2022) as well as levels of teacher retention (Allen et al., 2024; Perryman & Calvert, 2020). Thus, the role of AI applications in partially alleviating contributing factors should not be ignored.

If future research using AI assessment tools does not observe a consistently strong relationship between L2 phonological decoding and L2 reading outcomes, this has implications for education policy. For example, in such a scenario, a relaxed approach to requirements regarding teaching foreign language reading could be considered, especially if such an eclectic approach is found to yield more positive outcomes than any sort of phonics-only or phonics-above-all initiative. In practice, this could also mean that the limited amount of time available to primary language teachers (Collen & Duff, 2024) can be directed towards materials that are most effective in achieving outcomes in the specific context such as encouraging self-efficacy and self-regulated reading strategies (Graham et al., 2020).

If on the other hand, a strong relationship between FL phonological decoding and FL literacy skills does come to be observed, apps of this kind are capable of far more than acting as stand-alone assessment tools and can be embedded as a central teaching material for FL classes. For example, the app developed for present study has the capacity to offer personalised feedback, provide tips for articulatory realisation, as well as carry out both formative and summative assessment. As the scoring systems of apps such as this evolve in the years to come, this individually tailored approach holds promise for YLs, especially when they have such varying linguistic profiles and when neither removing struggling individuals from a class nor a one-size-fits-all approach is the optimal solution (Courtney et al., 2017). Another important contribution will be the increased capacity of practitioners to monitor linguistic development longitudinally by reviewing the data summaries presented on in-app dashboards which outline individual and whole-class progression. Such dashboards offer unique insight, especially for assessing linguistic development in instances where progression is difficult to perceive because it is slow (Porter, 2020; Woore et al., 2018) or occurs in a non-linear fashion (De Bot et al., 2013). Practitioners could also choose to adapt their teaching in response to the information provided by the AI to support, for example, users who fall behind class averages or to ensure instruction provides sufficient challenge for those obtaining consistently high scores.

It should be noted that although there was a clear correlation between the app scores, which were calculated via comparisons with native speaker pronunciations, and human coding, which placed less importance on similarity to a native speakers, we do not consider AI scoring to be able to supplant human judgement entirely at this stage and strongly advise against interpreting these data as such. For instance, according to Fig. 1, of all the SSC tokens that were coded as a reasonable and intelligible form by the human coder, the app's in-built mean SSC score was 52.3 %, suggesting that for a specific sound, the scoring system either struggled to detect perceptually accurate sounds or simply reserved its highest scores only for more nativelike productions by users, as may be expected given the way in which its scoring system works. For all SSC pronunciations deemed unintelligible in the human auditory coding (that is, for all sound tokens coded as 0 by the expert coder), the app's in-built mean SSC score was 46.5 %, suggesting stricter criteria may be required to ensure low scores are given for perceptually inaccurate tokens. Alternatively, AI research could look to automate assessment of pronunciation based on comprehensibility features (see e.g. Saito et al., 2022). Given that the scoring method employed by the app and the human coder differ in the present study, the app's scores could be validated more fairly by human assessors who compare the user utterances to recorded native speaker productions and provide perceived similarity scores out of 100. This would ensure that a) the app scores and the human coding is on the same percentage scale (rather than various combinations of binary, Likert and percentage scales) and b) that the same construct (native similarity) is being measured by both the app and the humans. Nevertheless, we argue that using the current auditory coding scales for both the SSC analysis and the word-level analysis aligns more closely with the advances in pronunciation research which have encouraged learners and practitioners to move away from aiming for nativelike forms (Levis, 2018). We maintain that with increased sensitivity, in combination with increasingly more advanced ASR models, the methodology explored here may well lead to replacement of human-led assessment in the future. At this stage, however, we recommend its use only as an informative guide into FL phonological decoding performance that can be delivered at scale to the benefit of both researchers and practitioners alike. Given that apps such as this are also capable of providing AI feedback and a high degree of interactivity, it is also important to weigh up the potential for increased learner engagement along with a greater amount of time dedicated to FL speaking against any discrepancy between human raters and automated measures.

6. Conclusion

This research proposes the use of apps that leverage AI models to assess phonological decoding among YLs of an FL. The scores in the app produced for the present research were found to significantly correlate with auditory coding by an expert human rater, both for individual symbol-sound correspondences and for the comprehensibility of words. This assessment tool is a methodological innovation for FL research, allowing the investigation of important questions for the field such as the strength of the link between phonological decoding, FL literacy skills, FL pronunciation and FL processing. In such investigations, the app output metrics should perform relatively consistently across contexts given that the relationship between the app scores and the auditory coding was not modulated by language, linguistic proficiency or recording quality. In summary, this study is an important first step in the validation of a tool that will prove vital for a variety of follow-on studies across a range of languages and beyond the UK. We now turn to some potential directions for such research.

Firstly, the present study indicated that the app's word level scores corresponded to how difficult the word was to understand. This is important because it allows the field to move away from native speaker models in the knowledge that FL learners will rarely if ever acquire nativelike pronunciation (Flege & Bohn, 2021; Levis, 2018; Saito, 2021). Nevertheless, one limitation of the current study is that the auditory coding was conducted by one individual which may bring into question how far the human auditory results generalise to other listeners. Future research should look to collect ratings from a variety of listeners (fellow L2 learners, native speakers of different varieties and different experiences of perceiving non-native speech) to determine how closely the app scores correspond to the ratings of these different groups. Such research would empower learners and practitioners to use the app in the knowledge that the scores correspond to ratings by the appropriate FL target for their specific contexts, especially if native speakers are unlikely to be the intended audience.

Secondly, it is worth noting that in the present study, the expert rater was familiar with the target items in advance which helped to ensure lexical biases played less of a role in auditory coding but may have affected the authenticity of the comprehensibility ratings. One consequential limitation was that intelligibility (whether the items were recognised irrespective of the difficulty in reaching that outcome) could not be assessed. Future research would do well to present items to raters who have yet to see or hear the items in order to establish whether or not the correct words are transcribed as a metric of intelligibility and to compare these values against the AI-generated app scores. However, any such designs would need to carefully control for factors affecting word recognition such as word frequency and phonological density, as well as item familiarity effects created by presenting the same item to the same rater on multiple occasions.

In short, exciting opportunities for further research and classroom practice lie ahead and the affordances that AI provides will undoubtedly serve to propel FL phonological decoding assessment forward.

CRediT authorship contribution statement

James Turner: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Alison Porter:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition. **Suzanne Graham:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition. **Travis Ralph-Donaldson:** Software, Resources, Methodology, Funding acquisition, Data curation, Conceptualization. **Heike Krüsemann:** Project administration. **Pengchong Zhang:** Project administration, Funding acquisition. **Kate Borthwick:** Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge funding from the Economic and Social Research Council (ESRC) (Grant Number: ES/X002500/1). We would also like to thank Niter LTD for developing a bespoke version of this app for the purposes of this research as well as the young language learners and their teachers who took part in the *Digital Empowerment in Language Teaching* (DELTEA) project, without whom this research would not have been possible. Finally, we would like to thank the anonymous reviewers and editors at *Research Methods in Applied Linguistics* for their useful comments which greatly helped to improve this manuscript. Any remaining errors are our own.

Appendix A. Pre-test walkthrough and avatar creation

Fig. 3 shows a screenshot from the app taken just after users were asked to read aloud an item from the second subsection of test items (*quime*). In the centre of the image, the graphic shows the end result of the app's real-time intonation tracking, amplitude recording, and segmentation mechanism which, on this occasion, finds 3 sounds. This also confirms that the app registered and recorded the speech. To the left, there is a wheel which shows the options for avatar customisation (e.g. body, eyes, mouth, hair, accessories). The avatar customisations that are unlocked also document the subsection of the task that the learner has reached (the second subsection in the present image). On the right, is the user's personalised avatar that takes form over the course of the task.

For comparison, Fig. 4 shows a screenshot just after reading a word from the third subsection of items (*abondent*). The learner now has the opportunity to customise the mouth from a range of options.

Narrative instructions in English were given to the learner throughout the course of the pre-test explaining how the task would develop, e.g. that there would be six sets of avatar features to unlock the more test items that were read aloud. Each subsection of the task was scaffolded with written (not auditory) prompts.

French Items		
Variable	**Item**	**Approximate Translation**
<qu> – [k]	coquin	cheeky
	quime	(Non-word)
<r> – [ʁ]	raide	straight
	rique	(Non-word)
<ch> – [ʃ]	déchets	rubbish
	chègue	(Non-word)
<u> – [y]	astuce	cunning
	ude	(Non-word)
<ou> – [u]	langouste	spiny lobster
	oude	(Non-word)
<on> – [ɔ̃]	abondent	abundance of
	onve	(Non-word)
<j> – [ʒ]	soja	soy
	jaude	(Non-word)

Appendix B. Test items used in the phonological decoding RAT

Spanish Items		
Variable	**Item**	**Approximate Translation**
<v> – [b]	visón	mink
	vem	(Non-word)
<qu> – [k]	peque	little one
	quim	(Non-word)
<ñ> – [ɲ]	leña	firewood
	ñol	(Non-word)
<r> – [r]	ramal	strand
	ril	(Non-word)
<j> – [x]	fijo	fixed
	jem	(Non-word)
<u> – [u]	lapsus	lapse
	ud	(Non-word)
<z> – [θ]	lazo	bow
	zal	(Non-word)

Appendix C. SSC Analysis: Pairwise comparisons of estimated marginal means for the app's post-hoc SSC score * SSC Variable interaction

Difference in SSC Auditory Coding scores between tokens coded as "1" in the app's post-hoc SSC score and tokens coded as "0" in the app's post-hoc SSC score

Variable	estimate	SE	*z* ratio	*p* value	significance
<j> – [x] (ES)	2.99	0.60	4.96	< 0.001	***
<z> – [θ] (ES)	1.78	0.48	3.68	< 0.001	***
<r> – [r] (ES)	3.05	0.90	3.37	< 0.001	***
<u> – [u] (ES)	2.20	0.97	2.27	0.02	*
<ñ> – [ɲ] (ES)	19.23	51.32	0.37	0.71	
<qu> – [k] (ES)	0.11	0.60	0.19	0.85	
<v> – [b] (ES)	−14.91	66.35	−0.22	0.82	
<ch> – [ʃ] (FR)	2.40	0.46	5.21	< 0.001	***
<j> – [ʒ] (FR)	2.55	0.64	3.97	< 0.001	***
<ou> – [u] (FR)	3.46	0.90	3.85	< 0.001	***
<u> – [y] (FR)	2.58	1.59	1.62	0.11	
<qu> – [k] (FR)	0.26	0.52	0.50	0.62	

Appendix D. Comprehensibility Analysis: Pairwise comparisons of estimated marginal trends for app's in-built SSC score * SSC Variable interaction

Estimates indicate for each individual SSC, the relationship between the app's inbuilt SSC score and the human auditory comprehensibility coding for the word as a whole

Variable	**estimate**	**SE**	***z* ratio**	***p* value**	**significance**
<j> – [x] (ES)	0.58	0.62	0.94	0.35	
<z> – [θ] (ES)	0.73	0.50	1.46	0.15	
<r> – [r] (ES)	1.46	0.87	1.69	0.09	.
<u> – [u] (ES)	0.61	0.51	1.19	0.23	
<ñ> – [ɲ] (ES)	0.04	0.66	0.07	0.95	
<qu> – [k] (ES)	0.08	0.84	0.10	0.92	
<v> – [b] (ES)	0.04	0.57	0.08	0.94	
<ch> – [ʃ] (FR)	1.56	0.65	2.39	0.02	*
<j> – [ʒ] (FR)	1.60	0.62	2.57	0.01	*
<ou> – [u] (FR)	0.37	0.56	0.66	0.51	
<on> – [ɔ̃] (FR)	0.75	0.72	1.04	0.30	
<u> – [y] (FR)	0.61	0.53	1.16	0.25	
<qu> – [k] (FR)	−2.01	0.77	−2.60	0.01	*
<r> – [ʁ] (FR)	−0.78	1.09	−0.72	0.47	

Appendix E. Comprehensibility Analysis: Pairwise comparisons of estimated marginal means for the app's post-hoc SSC score* SSC Variable interaction

Difference in modelled comprehensibility ratings between tokens coded as "1" in the app's post-hoc SSC score and tokens coded as "0" in the app's post-hoc SSC score

Variable	estimate	SE	*z* ratio	*p* value	significance
<j> – [x] (ES)	2.49	0.60	4.15	< 0.001	***
<z> – [θ] (ES)	1.24	0.55	2.24	0.03	*
<r> – [r] (ES)	1.73	1.05	1.65	0.10	.
<u> – [u] (ES)	1.19	1.07	1.12	0.26	
<ñ> – [ɲ] (ES)	22.73	289.63	0.08	0.94	
<qu> – [k] (ES)	0.74	0.73	1.02	0.31	
<v> – [b] (ES)	−1.10	0.93	−1.18	0.24	
<ch> – [ʃ] (FR)	2.11	0.65	3.27	< 0.01	**
<j> – [ʒ] (FR)	1.71	0.77	2.22	0.03	*
<ou> – [u] (FR)	0.18	0.77	0.23	0.82	
<u> – [y] (FR)	−0.66	1.91	−0.35	0.73	
<qu> – [k] (FR)	1.34	0.59	2.26	0.02	*

References

Allen, B., Ford, I., Hallahan, G., & Hannay, T. (2024). Teacher recruitment and retention in 2024: An exploration of recruitment challenges in disadvantaged schools [report]. https://www.gatsby.org.uk/uploads/education/2024-06-13-teacher-tapp-final-teacher-recruitment-and-retention-in-20241.pdf [Accessed 18th August 2025].

Apel, K. (2011). What is orthographic knowledge? *Language, Speech, and Hearing Services in Schools, 42*(4), 592–603. https://doi.org/10.1044/0161-1461(2011/10-0085)

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bassetti, B. (2024). Orthographic effects in the phonetics and phonology of second language learners and users. In M. Amengual (Ed.), *The Cambridge handbook of bilingual phonetics and phonology* (pp. 699–720). Cambridge University Press. https://doi.org/10.1017/9781009105767.032.

Bassetti, B., Escudero, P., & Hayes-Harb, R. (2015). Second language phonology at the interface between acoustic and orthographic input. *Applied Psycholinguistics, 36* (1), 1–6. https://doi.org/10.1017/S0142716414000393

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication, 49*(10–11), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro, & O.-S. Bohn (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). John Benjamins.

Boersma, P., & Weenink, D. (2024). Praat: Doing phonetics by computer [computer programme]. Version 6.1.28. https://www.praat.org.

Cable, C., Driscoll, P., Mitchell, R., Sing, S., Cremin, T., Earl, J., Eyres, I., Holmes, B., Martin, C., & Heins, B. (2010). Primary modern languages: A longitudinal study of language learning at Key Stage 2 [report]. https://eprints.soton.ac.uk/143157/1/DCSF-RR198.pdf [Accessed 18th August 2025].

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest, 19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Christensen, R. H. B. (2023). ordinal—Regression models for ordinal data. V2023.12-4.1 [R package]. https://cran.r-project.org/package=ordinal.

Chu, M. C., & Chen, S. H. (2014). Comparison of the effects of two phonics training programs on L2 word reading. *Psychological Reports, 114*(1), 272–291. https://doi.org/10.2466/28.10.PR0.114k17w0

Clark, M. M. (2017). *Reading the evidence: Synthetic phonics and literacy learning*. Witley Press.

Collen, I., & Duff, J. (2024). *Language Trends England 2024: Language teaching in primary, secondary and independent schools in England*. https://www.britishcouncil.org/sites/default/files/language_trend_england_2024.pdf [Accessed 18th August 2025].

Coltheart, M. (2005). Modeling reading: The dual-route approach. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 6–23). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch1.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100*(4), 589–608. https://doi.org/10.1037/0033-295x.100.4.589

Cook, V. (2010). The relationship between first and second language acquisition revisited. In E. Macaro (Ed.), *The continuum companion to second language acquisition* (pp. 138–155). Bloomsbury.

Courtney, L., & Graham, S. (2019). "It's like having a test but in a fun way": Young learners' perceptions of a digital game-based assessment of early language learning. *Language Teaching for Young Learners, 1*(2), 161–186. https://doi.org/10.1075/ltyl.18009.cou

Courtney, L., Graham, S., Tonkyn, A., & Marinis, T. (2017). Individual differences in early language learning: A study of English learners of French. *Applied Linguistics, 38*(6), 824–847. https://doi.org/10.1093/applin/amv071

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica, 32*(2), 133–143.

Darnell, C. A., Solity, J. E., & Wall, H. (2017). Decoding the phonics screening check. *British Educational Research Journal, 43*(3), 505–527. https://doi.org/10.1002/berj.3269

De Bot, K., Lowie, W., Horne, S. L., & Verspoor, M. (2013). Dynamic Systems Theory as a comprehensive theory of second language development. In M. Mayo, M. Gutierrez-Mangado, & M. Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 199–220). John Benjamins.

de Jong, N. H. (2023). Assessing second language speaking proficiency. *Annual Review of Linguistics, 9*, 541–560. https://doi.org/10.1146/annurev-linguistics-030521

Dudley, A., & Marsden, E. (2024). The lexical content of high-stakes national exams in French, German, and Spanish in England. *Foreign Language Annals, 57*(2), 311–338. https://doi.org/10.1111/flan.12751

Egan, C., Oppenheim, G. M., Saville, C., Moll, K., & Jones, M. W. (2019). Bilinguals apply language-specific grain sizes during sentence reading. *Cognition, 193*, 1–11. https://doi.org/10.1016/j.cognition.2019.104018

Erler, L., & Macaro, E. (2011). Decoding ability in French as a foreign language and language learning motivation. *Modern Language Journal, 95*(4), 496–518. https://doi.org/10.1111/j.1540-4781.2011.01238.x

Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). *Second language speech learning* (pp. 3–83). Cambridge University Press.

Fletcher, J. M., Savage, R., & Vaughn, S. (2021). A commentary on Bowers (2020) and the role of phonics instruction in reading. *Educational Psychology Review, 33*(3), 1249–1274. https://doi.org/10.1007/s10648-020-09580-8

Geçkin, V., Kiziltaş, E., & Çinar, Ç. (2023). Assessing second-language academic writing: AI vs. Human raters. *Journal of Educational Technology and Online Learning, 6*(4), 1096–1108. https://doi.org/10.31681/jetol.1336599

Gibson, H., & England, J. (2016). The inclusion of pseudowords within the year one phonics 'Screening Check' in English primary schools. *Cambridge Journal of Education, 46*(4), 491–507. https://doi.org/10.1080/0305764X.2015.1067289

Gimenes, M., Perret, C., & New, B. (2020). Lexique-Infra: Grapheme-phoneme, phoneme-grapheme regularity, consistency, and other sublexical statistics for 137,717 polysyllabic French words. *Behavior Research Methods, 52*(6), 2480–2488. https://doi.org/10.3758/s13428-020-01396-2

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. Remedial and special education. *Remedial and Special Education, 7*(1), 6–10. https://doi.org/10.1177/074193258600700104

Graham, S. (2020). *Assessment in the context of primary languages*. Research in Primary Languages (RiPL). https://ripl.uk/2020/06/02/assessment-in-the-context-of-primary-languages/ [Accessed 18th August 2025].

Graham, S., Woore, R., Porter, A., Courtney, L., & Savory, C. (2020). Navigating the challenges of L2 reading: Self-efficacy, self-regulatory reading strategies, and learner profiles. *Modern Language Journal, 104*(4), 693–714. https://doi.org/10.1111/modl.12670

Hamada, M., & Koda, K. (2008). Influence of first language orthographic experience on second language decoding and word learning. *Language Learning, 58*(1), 1–31. https://doi.org/10.1111/j.1467-9922.2007.00433.x

Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Studies in Applied Linguistics and TESOL, 16*(1). https://doi.org/10.7916/salt.v16i1.1261

Hayes-Harb, R., & Barrios, S. (2021). The influence of orthography in second language phonological acquisition. In *Language teaching, 54* pp. 297–326). Cambridge University Press. https://doi.org/10.1017/S0261444820000658

Hendrickson, K., Apfelbaum, K., Goodwin, C., Blomquist, C., Klein, K., & McMurray, B. (2022). The profile of real-time competition in spoken and written word recognition: More similar than different. *Quarterly Journal of Experimental Psychology, 75*(9), 1653–1673. https://doi.org/10.1177/17470218211056842

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*(1), 160–212. https://doi.org/10.1111/lang.12034

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2020). lmerTest: Tests in linear mixed effects models v3.1-3 [R Package]. https://cran.r-project.org/web/packages/lmerTest.

Lallier, M., & Carreiras, M. (2018). Cross-linguistic transfer in bilinguals reading in two alphabetic orthographies: The grain size accommodation hypothesis. *Psychonomic Bulletin and Review, 25*(1), 386–401. https://doi.org/10.3758/s13423-017-1273-0

Lampropoulos, G., Keramopoulos, E., Diamantaras, K., & Evangelidis, G. (2022). Augmented reality and gamification in education: A systematic literature review of research, applications, and empirical studies. *Applied Sciences, 12*(13), 1–43. https://doi.org/10.3390/app12136809

Landis, J. R., & Koch, G. G. (1977). The measurement of Observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, H., & Lee, J. H. (2024). The effects of AI-guided individualized language learning: A meta-analysis. *Language Learning & Technology, 28*(2), 134–162. https://hdl.handle.net/10125/73575.

Lenth, R., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Piaskowski, J., Riebl, H., & Singmann, H. (2024). Emmeans v1.10.4 [R Package]. https://www.cran.r-project.org/web/packages/emmeans.

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.

Liang, L., & Fryer, L. K. (2024). Phonological instruction in East Asian EFL learning: A scoping review. *System, 123*, 1–21. https://doi.org/10.1016/j.system.2024.103336

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly, 15*(3), 294–309. https://doi.org/10.1080/15434303.2018.1472265

McAuliffe, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. https://doi.org/10.21437/Interspeech.2017-1386

McAuliffe, M., & Sonderegger, M. (2022). MFA French and Spanish dictionaries v2.0.0. https://mfa-models.readthedocs.io/en/latest/dictionary [Accessed 18th August 2025].

McKay, P. (2005). *Assessing young language learners*. Cambridge University Press.

Morea, N., Kasprowicz, R. E., Morrison, A., & Silvestri, C. (2024). Diverse population, homogenous ability: The development of a new receptive vocabulary size test for young language learners in England using Rasch analysis. *Research Methods in Applied Linguistics, 3*(3). https://doi.org/10.1016/j.rmal.2024.100166

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice, 31*(3), 48–49. https://doi.org/10.1111/j.1745-3992.2012.00243.x

Nassaji, H. (2014). The role and importance of lower-level processes in second language reading. *Language Teaching, 47*(1), 1–37. https://doi.org/10.1017/S0261444813000396

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, and Computers, 36*(3), 516–524. https://doi.org/10.3758/BF03195598

Ng, S.-W. C. (2006). *The effects of direct instruction in phonological skills on L2 reading performance of Chinese learners of English*. University of London *[unpublished thesis]* https://core.ac.uk/download/pdf/111068931.pdf.

Perryman, J., & Calvert, G. (2020). What motivates people to teach, and why do they leave? Accountability, performativity and teacher retention. *British Journal of Educational Studies, 68*(1), 3–23. https://doi.org/10.1080/00071005.2019.1589417

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics, 2* (3), 1–8. https://doi.org/10.1016/j.rmal.2023.100083

Porter, A. (2020). Learning French sound/spelling links in English primary school classrooms. *EuroAmerican Journal of Applied Linguistics and Languages, 7*(1), 78–107. https://doi.org/10.21283/2376905x.11.187

R Core Team. (2024). *R: A language and environment for statistical computing [computer programme]*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org.

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly, 55*(3), 866–900. https://doi.org/10.1002/tesq.3027

Saito, K., Macmillan, K., Kachlicka, M., Kunihara, T., & Minematsu, N. (2022). Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition, 45*(1), 234–263. https://doi.org/10.1017/S0272263122000080

Sparks, R. L. (2021). Identification and characteristics of strong, average, and weak foreign language readers: The simple view of reading model. *Modern Language Journal, 105*(2), 507–525. https://doi.org/10.1111/modl.12711

Standards and Testing Agency. (2024). Phonics Screening Check. https://Www.Gov.Uk/Government/Publications/Phonics-Screening-Check-2024-Materials.

Unsworth, S., Persson, L., Prins, T., & De Bot, K. (2015). An investigation of factors affecting early foreign language learning in the Netherlands. *Applied Linguistics, 36* (5), 527–548. https://doi.org/10.1093/applin/amt052

van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology, 6*, 1–12. https://doi.org/10.3389/fpsyg.2015.01000

Woore, R. (2021). Teaching phonics in a second language. In E. Macaro, & R. Woore (Eds.), *Debates in second language education* (pp. 222–246). Routledge.

Woore, R. (2022). What can second language acquisition research tell us about the phonics 'pillar'? *Language Learning Journal, 50*(2), 172–185. https://doi.org/10.1080/09571736.2022.2045683

Woore, R., Graham, S., Courtney, L., Porter, A., & Savory, C. (2018). Foreign Language Education: Unlocking reading (FLEUR). A study into the teaching of reading to beginner learners of French in secondary school [report]. https://eprints.soton.ac.uk/445244/1/Foreign_Language_Education_Unlocking_Reading_FLEUR_A_study_into_the_teaching_of_reading_to_beginner_learners_of_Fre.pdf [Accessed 18th August 2025].

Young-Sholten, M., & Piske, T. (2008). Introduction. In T. Piske, & M. Young-Sholten (Eds.), *Input matters in SLA* (pp. 1–28). Multilingual Matters.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, (1), 131. https://doi.org/10.1037/0033-2909.131.1.3