Manuscript Running Header: Dimensions of Lexical Mastery and Their Relationships with Listening and Reading Proficiency.

Article Type: EMPIRICAL STUDY

Manuscript Title: Dimensions of Lexical Mastery and Their Relationships with Listening and Reading Proficiency Among Beginner-to-Low-Intermediate Learners of French

Author(s): Amber Dudley [a], Emma Marsden [b]

Author Affiliations: [a] University of York [b] University of Oxford.

Author notes / acknowledgements

CRediT author statement – **Amber Dudley**: conceptualization; methodology; data curation; investigation; formal analysis; visualization; project administration; writing – original draft; writing – review & editing. **Emma Marsden**: conceptualization; methodology; formal analysis; supervision; funding acquisition; project administration; writing review & editing.

Correspondence concerning this article should be addressed to Amber Dudley, University of York, Department of Education, University of York, York, YO10 5DD, amber.dudley@york.ac.uk.

Dimensions of lexical mastery and their relationships with listening and reading proficiency among beginner-to-low-intermediate learners of French

Abstract

The extent to which lexical mastery consists of interrelated types of lexical knowledge has been widely discussed. Yet, few studies have explored whether the lexical dimensions of knowledge and processing are relatively independent or their contribution to listening and reading, particularly among beginner-to-low-intermediate learners of languages other than English. The current study examines the extent to which lexical knowledge and processing are (a) relatively independent and (b) predict listening and reading. 218 beginner-to-low intermediate adolescent learners of French completed a battery of tests, yielding measures of high-frequency lexical knowledge (meaning-recognition, form-recognition, and form-recall), lexical processing (speed, stability, and automaticity of word recognition), and listening and reading proficiency. Results suggested that lexical knowledge and processing were relatively independent at this stage of proficiency. Lexical knowledge (but not processing) was found to strongly predict listening and reading. These findings underscore the importance of high-frequency vocabulary at these early stages.

Introduction

This study addresses two closely connected agendas. The first investigates the extent to which lexical knowledge and lexical processing fit within a componential view of lexical mastery.[1] Lexical processing has been conceptualised as a *potential* dimension of lexical mastery, together with the size and depth of lexical knowledge (Godfroid, 2020; Harrington, 2018). However, most empirical studies (e.g., Harrington, 2018; Pellicer-Sánchez & Schmitt, 2012) investigating relations between lexical knowledge and processing have focused on

intermediate and advanced learners, with, to the best of our knowledge, little attention paid to beginner learners. These issues thus warrant further investigation, particularly among low-proficiency learners when lexical representations are being established. The second relates to the relative and collective contribution of individual types of lexical knowledge and processing to listening and reading comprehension (In'nami et al., 2022; Jeon & Yamashita, 2022; Zhang & Zhang, 2022).

Henceforth, we refer to lexical knowledge and lexical processing as *dimensions* of lexical mastery and individual measures as *types* of lexical knowledge (e.g., form-recall) and lexical processing (e.g., mean response times [RTs]). Conceptualising lexical knowledge and processing as *dimensions* highlights that lexical mastery entails not only word knowledge but also the application of that knowledge during real-time language use (processing), whereas referring to specific *types* of measures reflects how these dimensions can be operationalised. This distinction in terminology broadly reflects Godfroid's (2020) expansion of Nation's (2023) framework, in which aspects of lexical knowledge (form, meaning, and use) are operationalised in terms of offline and online measures that capture controlled, explicit word knowledge and automatized explicit and implicit lexical knowledge, respectively.

The aim of the current study is to explore (a) the extent to which lexical knowledge and processing represent relatively independent constructs and (b) the extent to which types of these two dimensions predict L2 proficiency in listening and reading. 218 beginner-to-low-intermediate English-speaking classroom learners of L2 French completed three vocabulary tasks comprising measures of lexical knowledge: form-recall (English-to-French translation), meaning-recognition (word–definition matching task), and form-recognition (lexical-decision) tasks, each containing the same 50 high-frequency words sampled from learners' school curriculum. The lexical-decision also provided two lexical processing measures (RTs indexing speed of word recognition and coefficients of variation [CVs] indexing stability and

automaticity of word recognition). The four proficiency measures were the listening and reading components from one low-stakes (the *diplôme d'études en langue française* [DELF; Diploma in French Language Studies]) proficiency test and one high-stakes (the General Certificate in Secondary Education [GCSE]) proficiency test.

Background Literature

Lexical knowledge as a theoretical concept

It is generally accepted that lexical knowledge is made up of multiple components. Nation's (2013) framework of lexical knowledge identifies 18 components, covering receptive and productive mastery of word form, meaning, and use (see Table 1 in Appendix S1). This framework has since been updated by Godfroid (2020) to include a processing dimension, a point we return to later in the Background Literature. Although Nation's framework provided a comprehensive description of word knowledge, it did not aim to specify whether certain components are acquired before others or the extent to which they are interrelated (Nation, 2020).

Studies investigating lexical knowledge and interactions of its multiple components are limited in number. The available studies, however, can be broadly divided into two sets (Yanagisawa & Webb, 2019). The first set explores the depth of lexical knowledge and focuses on the quality of *multiple* components of lexical knowledge. González-Fernández and Schmitt (2020), for instance, tested 144 beginner-to-advanced Spanish-speaking learners' recognition and recall knowledge of 20 words in English across four components: the form-meaning link, derivatives, multiple meanings, and collocations. Structural equation modelling showed that each component significantly contributed to the global construct of lexical knowledge, with factor loadings exceeding .81, suggesting a unidimensional conceptualisation best characterised L2 lexical knowledge. In a replication of this study with

314 beginner-to-advanced learners of English from two first language (L1) backgrounds (Chinese *n* = 170; Spanish *n* = 144), González-Fernández (2022) reported that L2 lexical knowledge served as a unidimensional construct, regardless of learners' L1 background. The second set—the focus of our study—examines the strength of lexical knowledge and compares different types within a *single* component (e.g., the form-meaning link). These types include form-recall (e.g., via L1-to-L2 translation), meaning-recall (e.g., via L2-to-L1 translation), form-recognition (e.g., via Yes-No tests), and meaning-recognition (e.g., via word–definition matching tasks). Existing research (e.g., Laufer & Goldstein, 2004; Webb, 2005) suggests that recognition and recall are interrelated and that L2 learners typically achieve higher scores in recognition than recall measures. Laufer and Paribakht (1998), for instance, reported that among 182 L2 learners of English of different proficiencies, recognition (where responses were selected from options) and recall (where responses were produced without options provided) scores were strongly correlated (*r* = .72-.89), but recognition scores always exceeded recall scores, especially with low(er)-frequency words. Due to practicalities, the number of items in these studies has tended to be small, ranging from 10 (Chen & Truscott, 2010; Webb, 2005) to 20 (González-Fernández, 2022; González-Fernández & Schmitt, 2020) words. Recent recommendations that at least 30 items are needed to draw generalisations (Gyllstad et al., 2021) thus provide additional rationale for further research.

Another motivation for further research is the almost exclusive focus on testing *intermediate and advanced* learners of L2 *English*. In a recent meta-analysis exploring the relationship between lexical knowledge and L2 listening and reading, Zhang and Zhang (2022) reported that more than 80% of the 100 + individual studies examined English as the L2. Testing beginner-to-low-intermediate learners in languages other than English is especially important for understanding how words are acquired, stored, and represented in the

mental lexicon in the early stages of L2 development. It has been argued (see, e.g., Bordag et al.'s [2022] Ontogenesis Model of the L2 Lexical Representation) that the encodings of lexical representations are often 'fuzzy' (i.e., inexact or ambiguous) during these early stages. For instance, an L2 learner may recognise a given word form but not know its meaning, or they may alternate between two (or more) word forms for a particular meaning. However, given the paucity of research among beginner-to-low-intermediate learners, especially in low-exposure, instructed contexts, it is not yet clear the extent to which this fuzziness affects the relations between multiple types of lexical knowledge and, perhaps, lexical processing, a point we return to later in the Background Literature.

Another factor that can further increase the extent of fuzziness in lexical representations is the degree of orthographic depth in the target language (see, e.g., Schmalz et al., 2015). In orthographically shallow languages (e.g., Spanish), most phonemes map onto a single grapheme, and most graphemes map onto a single phoneme, at least within one variety of the language. Lexical development typically occurs at a much faster rate in these languages than in orthographically deeper languages (Carrillo et al., 2013), like French, where a phoneme can map onto multiple graphemes. However, there is transparency in the other direction in French, as one grapheme generally maps onto a single phoneme. In English, there is even greater opaqueness, with multiple graphemes mapping onto multiple phonemes and multiple phonemes mapping onto multiple graphemes. Although our study does not manipulate orthographic transparency, it does respond to the need to expand the evidence-base about lexical knowledge and processing to include languages with different orthographic depths to English.

Relations between lexical knowledge and proficiency in listening and reading

Recent meta-analyses have generally reported that lexical knowledge strongly correlates with L2 listening (r = .56, 95% CI [.51, .61]; In'nami et al., 2022) and reading (r = .72, 95% CI [.64, .79]; Jeon & Yamashita, 2022). Considerable variation in correlations, however, has been observed for both reading (between .08 and .95) and listening (between .13 and .85; Zhang & Zhang, 2022).

One reason for this variation may be the types of lexical knowledge tested in the individual studies. Zhang and Zhang's (2022) meta-analysis found that meaning-recall (r = .66, 95% CI [.58, .71]) correlated more strongly with reading comprehension, followed by form-recall (r = .55, 95% CI [.48, .63]) and meaning-recognition (r = .53, 95% CI [.49, .57]). Caution, however, must be exercised when interpreting these comparisons, given the largely overlapping confidence intervals of the three r coefficients.

Although meaning-recall typically predicts L2 reading the most strongly, a different tendency has been reported for listening. Zhang and Zhang's (2022) meta-analysis of 47 listening studies found that form-recall (r = .63, 95% CI [.53, .72]) correlated more strongly with listening than meaning-recall (r =.58, 95% CI [.54, .62]) or meaning-recognition (r = .50, 95% CI [.41, .58]). Again, overlapping confidence intervals between all three types of knowledge suggest that this order may not be entirely reliable.

Another factor contributing to the variability in correlations between different studies may be the modality of lexical knowledge tested. Zhang and Zhang (2022) reported that the relationship was significantly stronger between written lexical knowledge and reading (r = .60, 95% CI [.54, .63]) than between aural lexical knowledge and reading (r = .49, 95% CI [.46, .54]), with only negligibly overlapping CIs. In contrast, no significant difference between the degree of association was found between aural lexical knowledge and listening (r = .60, 95% CI [.54, .65]) and between written lexical knowledge and listening (r = .52,

95% CI [.44, .59]). Similarly, there was substantial overlap in the CIs around the correlations between written lexical knowledge and reading (r = .60, 95% CI [.54, .63]) and written lexical knowledge and listening (r = .52, 95% CI [.44, .59]).

An important question arising from this research is the extent to which the predictive relationships between lexical knowledge and listening and reading comprehension vary as a function of proficiency. Given current evidence, such a question is difficult to answer, as most research has focused on intermediate and advanced learners. Stanovich (2000) proposed that a reciprocal causal relation exists between lexical knowledge and reading, such that as lexical knowledge improves, so does reading proficiency, and, equally, as reading proficiency improves, so does lexical knowledge. Thus, it may be that in the early stages of learning, relations between lexical knowledge and reading (and potentially also listening) proficiency are not as strong as has been found in research with more proficient learners.

Taken together, these findings suggest that the relative importance of different types of lexical knowledge for language comprehension may depend—to some extent—on the type of knowledge assessed and the modality in which it is assessed, though findings are mixed. Further research is therefore needed to explore the extent to which these findings pertain in languages other than English and among lower proficiencies, where lexical mastery is emergent, and some knowledge types may be more robust than others.

Is lexical processing a component of lexical mastery?

When listening or reading in an L2, learners need to be able to retrieve words quickly, automatically, and accurately (Just & Carpenter, 1992). Remarkably little research, however, has investigated relations between lexical knowledge and processing, especially among *beginner-to-low-intermediate* L2 learners. As such, the extent to which lexical processing—

as indexed by the speed, stability, and automaticity of word recognition—relates to lexical knowledge remains unclear, despite calls in the literature to investigate this further.

These calls include proposals from Harrington (2018) and Godfroid (2020) to expand existing multi-componential frameworks of lexical knowledge, such as Nation's (2013), to include a processing dimension. For example, Harrington (2018, p. 67), in his Lexical Facility proposal, argued that vocabulary size and processing skill could be complementary and, critically, interrelated indices of L2 lexical mastery, suggesting that when these indices are combined (as a composite measure), they may provide "a more sensitive measure of individual differences in L2 vocabulary than vocabulary size alone". Similarly, Godfroid (2020) added automaticity to Nation's (2013) framework while also illustrating how each aspect of lexical knowledge could be measured both offline and online (see Table 2 in Appendix S1 for the expanded framework).

Given our current understanding of how words are stored and represented in the mental lexicon, the degree of interrelatedness between lexical knowledge and processing may vary as a function of proficiency. As discussed previously, the encodings of lexical representations are often inexact or ambiguous during the early stages of acquisition, and this fuzziness may influence the accuracy and efficiency of word recognition. For instance, Perfetti's (2007) Lexical Quality Hypothesis predicts that processing stability—a hallmark of fluent listening and reading (Segalowitz, 2010)—develops as a function of lexical knowledge. That is, when lexical representations are inexact or ambiguous, processing stability may not be expected.

Studies investigating these proposals are limited in number. (e.g., Harrington, 2018; Pellicer-Sánchez & Schmitt, 2012). In general, the available studies have reported negative correlations between vocabulary size and mean RT (i.e., *more* vocabulary is associated with *faster* responses) among intermediate and advanced learners, thus supporting the proposal

that vocabulary size and processing skills might be related. However, no correlation between the two measures was found in another study (Miralpeix & Meara, 2010, as cited by Harrington, 2018) when intermediate and advanced L2 English students were tested using a separate vocabulary size test and lexical-decision task. Harrington suggested that Miralpeix and Meara's lack of correlation was due to the relative ease of the target words: The mean RT was 815ms among their learners, only 31ms slower than the native speakers in their study.

However, vocabulary size is only one index of vocabulary knowledge, with depth of knowledge being another key aspect. During the early stages of L2 development, an empirical question arises as to whether one can reasonably expect beginner-to-low-intermediate learners to demonstrate processing stability when their vocabulary sizes are relatively small, and the depth of their knowledge of words extracted from their specific learning context could be variable, potentially depending on the type of measure used (e.g., form- or meaning-recognition or recall) to assess it. As far as we know, no study has examined this question to date.

The lack of research is partly due to the use of explicit–declarative measures over more implicit–procedural measures to test L2 lexical knowledge. However, Godfroid (2020, p. 433) suggests that explicit–declarative measures may only represent the "tip of the iceberg". Lexical-decision (Yes–No RT) tasks are one of several implicit–procedural measures that can provide information about how quickly, automatically, and accurately learners can access their knowledge (Godfroid, 2020). In these tasks, participants are presented with individual words and pseudowords (non-words) and must press a button to indicate whether they think the word is real or not. These tasks can thus elicit both explicit–declarative (e.g., form-recognition) and implicit–procedural (e.g., RT) measures.

Hui et al. (2025) recently highlighted the importance of incorporating RT measures when assessing lexical knowledge. They investigated the extent to which *response time* tasks

(i.e., a Yes–No RT test [affording both accuracy and RT measures] and a masked repetition priming task) tapped into qualitatively different dimensions of word knowledge to *untimed, accuracy-based* tasks (i.e., untimed meaning recognition test and form recall tests) among 145 L2 learners of English. Confirmatory factor analyses revealed that these measures could be explained by one or two psychometric dimensions but that, ultimately, the one-factor solution, where the aforementioned measures reflected one underlying construct, was preferred due to its simplicity.

It is important to acknowledge, though, that mean RT is only one indicator of processing efficiency. From RT data in lexical-decision tasks, a 'coefficient of variation' (CV; Segalowitz et al., 1998) can be extracted. The CV is the ratio between the amount of variation (standard deviation) in RT for each participant (across items) and the mean RT for each participant (across items). A smaller coefficient indicates that participants can respond repeatedly at similar speeds, suggesting they rely (more) on automatic rather than controlled processing. The CV is typically considered a measure of processing stability. However, it can also index processing automaticity. Although, intuitively, automaticity is reflected in faster processing (e.g., Hulstijn et al., 2009), faster processing does not necessarily indicate automaticity. Automaticity is more than simply faster processing; it is "the observable speed, accuracy, and fluidity of skill execution" (Segalowitz et al., 1998, p. 53). As such, RT and the CV must correlate within any given sample of learners for the CV to be interpreted as an index of automaticity.

There are also instances where the CV has been found to increase initially in a learner's developmental trajectory, contrary to expectations that it would go down (e.g., Hui, 2020; McManus & Marsden, 2019; Solovyeva & DeKeyser, 2018). Hui (2020), for instance, investigated the trajectory of CV (as an index for processing variability) in the early stages of word learning in both intentional and incidental conditions among 35 native English speakers

learning 16 Swahili–English word pairs. In this study, Hui found a roughly inverted U-shaped trajectory in the CV in the intentional condition. That is, processing stability appeared to increase during the early stages when declarative and procedural lexical knowledge was being acquired. However, it was only once the words had been learnt (as indicated by the peak in accuracy) that processing stability improved, with CV starting to decrease, consistent with automatization.

Despite the aforementioned theorisation, few studies have compared the contributions to listening and reading comprehension of speed, stability, and/or automaticity of word recognition relative to lexical knowledge, such as recognition and recall. One study by van Gelderen et al. (2004) found that among 397 13-to-14-year-old Dutch-speaking learners of English, lexical and metacognitive knowledge (β = .26 and .70, respectively) —but not word recognition speed—significantly predicted reading comprehension. Similarly, Alshehri and Zhang (2022) reported that lexical knowledge (R^2 = .414)—but not word recognition speed (R^2 = .034)—significantly moderated L2 reading comprehension among 220 Arabic-speaking intermediate-to-advanced learners of English in a Saudi university.

In terms of listening comprehension, Andringa et al. (2012) tested 113 L2 Dutch speakers. Although linguistic knowledge (i.e., vocabulary [.68], grammatical processing accuracy [.77], and segmentation accuracy [.64]) and processing speed (i.e., semantic processing [–.61], grammatical processing [–.41], and segmentation speed [–.36]) showed moderate-to-strong correlations with listening, processing speed did not contribute any unique variance (in addition to linguistic knowledge) in the structural equation models.

Larger vocabulary sizes and faster RTs may also be more strongly associated with better comprehension when comprehension is relatively robust in the first place and, in turn, possibly associated with proficiency. For example, Hui and Godfroid (2021) tested the role of processing speed and automaticity in L2 listening among 44 low-intermediate-to-advanced

Chinese learners of English. In the mediation (path) analysis, they found that the total effects of spoken vocabulary size (β = .58, p < .001) and lexical processing speed (β = –.50, p = .001), as indexed by accuracy and RT in an auditory Yes–No test, respectively, were stronger than the direct effects alone (vocabulary size: β = .35, p = .07; processing speed: β = –.39, p = .01), when a measure of formulation of propositional meaning was specified as the mediator. These findings thus indicate that the importance of vocabulary size and lexical processing speed for comprehension may depend on propositional comprehension.

Evidence that a lexical knowledge measure incorporating an automatization construct can be a strong predictor of listening proficiency has been provided by Saito et al. (2023). They investigated individual differences in listening proficiency among 126 beginner-to-advanced Japanese learners of English. They reported that "phonological lexical knowledge" explained 77.60% of the variance in the full regression model (R^2 = .507), whereas perceptual, cognitive, and metacognitive factors only contributed 0.40% to 21.30%. Of the phonological lexical knowledge variables, automatization—as measured by both accuracy and fluency (CV) scores on a lexicosemantic judgment task—was the strongest predictor (55.30%), followed by phonologization (20.80%) and then generalization (1.50%). Importantly, of the two automatization measures, accuracy, in fact, contributed 49.90%, but CV—the measure typically associated with automatization—contributed 5.40%, with correlation and factor analyses showing that accuracy and CV were relatively independent.

In sum, given the paucity of studies and mixed findings, further research is needed to explore (a) the extent to which types of lexical processing relate to types of lexical knowledge and (b) the relative importance of each type in predicting listening and reading comprehension.

The current study

The current study investigated the extent to which lexical knowledge (form-recall, form-recognition, and meaning-recognition) and processing (speed, stability, and automaticity of word recognition) relate to each other and, in turn, predict L2 listening and reading among 218 16-year-old classroom learners of French in England. It thus addresses two main research questions (RQs):

RQ1: To what extent are lexical knowledge and processing relatively independent of each other?

Perfetti's (2007) Lexical Quality Hypothesis argues that lexical processing stability and automaticity are by-products of *complete* lexical representations. Given the lack of research among learners with relatively small lexicons, it is not yet clear the extent to which lexical representations must be complete *across* the mental lexicon for processing stability to be observed. It may be that lexical knowledge of multiple types (e.g., form-recall, form-recognition, meaning-recognition) within a set of relatively familiar words is a sufficient condition for processing stability and automaticity within that set of words. As such, one might expect moderate-to-strong relationships between lexical knowledge and processing measures. Alternatively, there may be little-to-no relationship between lexical knowledge and processing measures, due to two opposing forces influencing the lexical processing measures, that is, the learning of new words, which are initially processed with less stability (increasing both mean RT and the CV) and the automatization of already established word representations (decreasing both mean RT and the CV).

RQ2: To what extent do lexical knowledge measures and processing measures predict listening and reading comprehension?

Based on previous research (e.g., Zhang & Zhang, 2022), we anticipated that each type of lexical knowledge would predict listening and reading. However, we were largely equipoised to the relative order of importance of each type, given the lack of clear differences (i.e., overlapping confidence intervals) between each type and our new population of learners. Given the paucity of available research, we were similarly equipoised to the extent to which processing measures (i.e., the speed, stability, and automaticity of word recognition) would determine listening and reading.

To avoid fatigue and attrition within and between sessions, we did not test meaning-recall in the main study, even though it is a known predictor of listening and reading comprehension. Following data collection in 2022, we were able to add a meaning-recall task to the study in 2023. Given that this task was taken by a subset of participants (n = 87) and tested a subset of the words (k = 30), the analyses of this dataset are not reported in the main results section and are mentioned only briefly in the Discussion to highlight their alignment with previous studies on the topic.

Method

Participants

Participants were 218 beginner-to-low-intermediate[2] English-speaking learners of French from 89 state-funded secondary schools across England, with a mean of 2.45 learners from each school. Of the 218 participants, 174 identified as female and 37 as male; eight preferred not to say or to self-describe. On average, participants reported learning French from 9.68 (95% CI [9.31, 10.06]) years of age.

At the time of testing, participants had recently (within the previous one-to-six weeks) completed their General Certificates in Secondary Education (GCSEs; national high-stakes exams taken by 15-to-16-year-olds in England) after about 400-450 hours of instruction in French. (For more information about the GCSE exams, see Appendix S2). Participation was optional, and learners were recruited via their school French teacher and paid £25 or £35 for their participation in two or three sessions, respectively. The University of York granted ethics approval.

Critical words

Fifty words were selected from a bank of high-frequency words (i.e., within the first 2,000 most frequent words in the French language, according to Lonsdale & Le Bras, 2009). The overarching principle guiding the selection of critical words was their presence in the curriculum. In this context, the curriculum was considered the GCSE wordlists published by the leading awarding organisations: AQA (2016) and Pearson Edexcel (2018). Although these wordlists were not compulsory, they have been heavily used by textbook writers (e.g., Hawkes & Lillington, 2016) and frequently used by teachers to guide lesson planning and exam preparation (see Marsden et al., 2023 for further discussion). These words, therefore, had the possibility of being familiar—to varying degrees—among our participants and were thus likely to provide variance in the different measures of knowledge and processing. Appendix S3 includes further information about the critical words.

Vocabulary Measures

Participants completed three written vocabulary tasks, yielding three lexical knowledge measures and two lexical processing measures. The tasks included tests of (a) form-recall, (b) meaning-recognition, and (c) lexical-decision (producing a form-recognition measure, RTs,

and CVs). The first two tasks were without a time constraint, and the third with (i.e., participants had to respond to each trial within 3,000ms).

The same 50 critical words were used in each test to ensure test equivalence between measures. Critical words were randomly mixed with distractors in the meaning-recognition task and pseudowords in the lexical-decision to reduce practice or priming effects. Only critical words—not distractors or pseudowords—were included in our analyses.

The form-recall task was completed before the meaning-recognition task in line with previous research (e.g., González-Fernández, 2022; González-Fernández & Schmitt, 2020), with the lexical-decision task being administered at least 24 hours after the first two tasks. The existence of any test effects (e.g., from priming) resulting from task order (such as the potential bolstering of any relationship between lexical knowledge, speed, stability, and automaticity) is addressed in the *Discussion*. At the same time, we argue that such effects are mostly unavoidable when testing different types of knowledge of the same set of words, and it is an empirical question as to how these effects can, or even need to, be mitigated.

We acknowledge that the exclusive use of written lexical knowledge measures is a limitation of our study, given the reduced face validity of using written measures to predict listening proficiency (McLean et al., 2024). As such, our battery of measures was arguably not comprehensive. Indeed, it is likely that aural measures of lexical knowledge could correlate more strongly with listening proficiency than written measures. For example, Zhang and Zhang's (2022) meta-analysis reported that aural lexical measures (r = .52, k = 30) had a numerically—though not significantly—stronger relationship with listening than written lexical measures (r = .60, k = 15). However, this does not reduce the value of exploring relationships between written measures of lexical knowledge and listening proficiency. Understanding the relationship between written word knowledge and listening remains important, particularly in instructed settings. For example, such findings may inform thinking

about the value of phonics (i.e., grapheme–phoneme correspondence) instruction. Additionally, on a practical level, we wanted to avoid fatigue and attrition within and between sessions, with a hard-to-reach population sacrificing vacation immediately after a series of between 18 and 30 high-stakes national exams. As Stewart et al. (2021, p. 57) observe, "it may be possible to identify the best trade-off in terms of practicality and efficacy in situations where researchers wish to measure proficiency on multiple skills but only have limited time to test vocabulary knowledge".

All materials, data, and analyses are available via our Open Science Framework repository (https://osf.io/hmb8y/?view_only=7e305ae7ae544679a225d9fa5a97f334).

Form-recall task

This task involved word-level L1-L2 translation, administered via Qualtrics (www.qualtrics.com). Participants were presented with a randomised list of the English translations of the 50 target words and asked to translate the words into French (see Figure 1). Participants were told that they could enter the word with or without a determiner ("the word for 'a' or 'the'") and were asked to include accents where necessary (an accent bar was provided). For instructions, see Appendix S4.

[Insert Figure 1 approximately here]

Form–meaning-recognition task

The Context-Aligned Two Thousand Test (CA-TTT), administered via Qualtrics, is a test of form–meaning-recognition developed by Dudley et al. (2024), drawing on Batista and Horst's (2016) *Test de la Taille du Vocabulaire.* The CA-TTT contains 120 target words (including the 50 critical words) and 120 distractors taken from the 2,000 most frequently

occurring words in French (Lonsdale & Le Bras, 2009). English definitions are presented in clusters of three, and participants must select which of the six French words (randomised between definitions and participants) best matches the definitions (see Figure 2).

[Insert Figure 2 approximately here]

Lexical-decision task

A lexical-decision task was administered to measure the speed, stability, and automaticity with which learners could retrieve (likely familiar) words from the mental lexicon (i.e., the form-recognition aspect of lexical processing). Alternatives, such as speeded matching and lexicosemantic judgment tasks to capture meaning-recognition, were considered inappropriate as they involve more than one process (e.g., considering alternative answer options before responding) and can thus only provide an indirect and coarse measure of lexical processing (Harrington, 2018; Hui & Jia, 2024).

The lexical-decision involved a speeded Yes–No vocabulary test administered via Gorilla (www.gorilla.sc). Test items were randomised within participants and included 60 target (real) words (including the 50 critical items) and 60 pseudoword distractors. Following previous research (e.g., Vandenberghe et al., 2021), pseudowords were created using the target (real) words: The first letter of each target word was replaced (consonants with consonants; vowels with vowels) while ensuring that the resulting non-word was phono-tactically legal.

Participants were presented with one word at a time in the centre of the screen and had to indicate whether the word was real or fake (left arrow key for real; right for fake). Participants were instructed to answer as quickly as possible while also maintaining accuracy. Each word was preceded by a fixation cross (for 250ms) with a pause of 100ms before and

after the fixation cross. The word remained on the screen for up to three seconds. If no key was pressed during that time, the next word appeared. After reading the instructions, participants completed a practice round of four items (two real words and two pseudowords) before starting the main task.

Measures of listening and reading proficiency

GCSE. Independently of our study, participants completed high-stakes national GCSE exams in French between May and June, just before our data collection. These exams comprised four skill-specific papers: listening, reading, writing, and speaking. For each paper, students were awarded a level (grade; between 1 and 9). In the August following our study, participants sent photographic evidence of their GCSE results, including their overall and skill-specific levels, of which only listening and reading were included in our study.

DELF. To assess comprehension, independent of performance in the high-stakes tests for which learners had been prepared, we administered the listening and reading components of the CEFR A2 Junior version of the *Diplôme d'études en langue française* (France Éducation International, n.d.). In these measures, participants listened to or read short passages of text and then answered multiple-choice questions. These measures were selected for two reasons: (a) they met all 17 of the Association of Language Testers in Europe's quality standards of test construction, administration and logistics, marking and grading, test analysis, and communication with stakeholders (ALTE, 2023) and (b) participants were not trained on them. The lack of training meant that the DELF listening and reading measures were likely to be a more accurate indication of learners' true abilities (in contrast to their test-taking abilities, as might be the case with the GCSE). As such, it was possible that relationships between listening and reading and lexical knowledge and processing would be more observable in the DELF than the GCSE data, as the GCSE data may (also) index general school performance and/or test-taking abilities.

Procedure

Participation was voluntary, with data collected between June and August 2022 and between June and August 2023.[3] The tests were administered online as part of a large study investigating language proficiency among French and Spanish students, involving two 90-minute sessions (followed by an optional 90-minute session for the larger study). Participants completed the first two sessions in their own time, without supervision[4], with the untimed tasks (form-recall then meaning-recognition) in the first session and, at least 24 hours later, the timed task (lexical-decision) and the DELF listening and reading comprehension tests in the second session.

Scoring

Scoring and marking criteria are in Appendix S5.

Accuracy data

Form-recall. Answers were scored on an integer scale from zero to two. Two points were given to answers matching the target (researcher-envisaged) translation and to any other correct translations of the English cue (e.g., if the cue were *to live*, the answer *vivre* would receive two points, even though the target answer was *habiter*). Misspelt but correct translations were scored either as zero or one, depending on the severity of the misspelling, as determined by the normalised Levenshtein distance between the target and answer, calculated using the *ratio()* function from the *python-Levenshtein* library (*python-Levenshtein*, n.d.). We used a normalised distance of .7 as a threshold, assigning one point to all answers with a score equal to or greater than the threshold value and zero to those falling below. No points were deducted for omissions or incorrect use of accents and/or determiners. Incorrect or blank answers were awarded zero points. Three independent raters scored answers. Interrater

agreement was very high (percentage agreement: 91%; Fleiss' Kappa: 90%, $p < .001$). Scores were thus averaged across the three raters and rounded to the nearest integer.

Meaning-recognition. Responses were scored as a binary outcome (1 for selecting the correct definition for the target word; 0 for incorrect selections).

Form-recognition. The accuracy percentage (i.e., the proportion of correct responses to the 50 critical words) from the lexical-decision was used to measure form-recognition. We used the index of signal detection theory (I_{SDT}; Huibregtse et al., 2002) to correct for false alarm rates (i.e., the proportion of incorrect 'YES' responses to pseudowords), as it: (a) considers individual patterns in guessing behaviour and (b) has been shown to correlate strongly with existing multiple-choice vocabulary size tests (Zhang et al., 2020). To calculate the I_{SDT}, we used the formula:

$$I_{SDT} = 1 - \frac{4\,(1-f) - 2(h-f)(1+h-f)}{4(1-f) - (h-f)(1+h-f)}$$

where h is the hit rate, and f is the false alarm rate.

DELF listening and reading. The answer key provided by the DELF test developer was used to score the listening and reading tasks: One point was awarded for each correct answer, with a maximum of 25 points per task. The score for each test was converted to a percentage.

RT and CV data

Only lexical-decision data for the 50 critical words were cleaned and analysed, and data from the 10 non-critical words were removed from the dataset prior to cleaning. To clean the remaining data, we removed RTs (a) from participants with false alarms rate of > .50, (b) faster than 300ms (after the onset of the word) and slower than 2,000ms (after the offset of

the word), (c) for incorrect 'No' responses (i.e., 'No' responses to real words), and (d) for the pseudoword distractors. Using the cleaned dataset, we calculated, for each participant, a mean RT and CV by dividing the standard deviation [SD] of their RTs by their mean RT (Segalowitz et al., 1998).

Analyses

Descriptive statistics, reliability indices, and Spearman's correlations were calculated to establish the degree of interrelatedness between each knowledge, processing, and proficiency measure.

To explore the extent to which our lexical knowledge and processing measures were relatively independent of each other (RQ1), we fitted a one-factor confirmatory factor analysis (CFA) model, using the *cfa()* function from the *lavaan* package (Rosseel, 2023). In this model, measures of lexical knowledge (form-recognition, form-recall, and meaning-recognition) and processing (mean RT and CV) loaded onto a single 'lexical mastery' factor. Given the non-normal distribution of the data (as indicated by Mardia's Multivariate Normality Test), we used the robust maximum likelihood estimator to compute the model. In addition, to account for method effects resulting from the non-independence of mean RT and CV, we allowed the residual terms of the two processing measures to covary (Brown, 2015).

We did not compare the one-factor model with a two-factor model, due to the difficulties associated with constructing a two-factor solution—comprising 'knowledge' and 'processing' factors—that could be meaningfully compared with a one-factor solution. Specifically, the 'processing' factor, with only two indicators (mean RT and CV), would have been just-identified, as long as no residual covariance between the indicators was estimated. However, our model required a residual covariance between the two processing measures, which introduced an additional parameter. This would have made the processing factor

under-identified, as the number of free parameters would have exceeded the number of distinct data points in the observed covariance matrix. This under-identification would have rendered the parameter estimates—and any subsequent interpretation—unreliable.

Instead, to address RQ1, we focused on assessing the fit of the one-factor solution, by inspecting both (a) the standardised root mean square residual (SRMR) and the comparative fit index (CFI) as global fit indices given our small sample size[5] and (b) factor loadings (i.e., standardised regression coefficients) as local fit indices. Given the flexibility of CFA to accommodate different model specifications, constraints, and estimation approaches, we invite and strongly encourage interested readers to conduct secondary analyses of our openly available dataset.

To investigate the collective and relative contribution of each knowledge and processing measure to listening and reading (RQ2), we calculated hierarchical (stepwise) regression models. For the (linear) DELF data, we used the *lm()* function in the R environment (R Development Core Team, 2014) and for the (ordinal) GCSE data, *polr()* from the MASS package (Ripley et al., 2022).

We first entered the overall mean accuracy percentages for the knowledge measures—form-recall, meaning-recognition, and form-recognition—as predictors, given the widely observed relations between lexical knowledge and listening and reading comprehension (Step 1). We then added mean RT *or* CV as a predictor (Step 2). Analyses of variance were conducted to compare the linear (DELF) models and likelihood ratio tests to compare the ordinal (GCSE) models. All variables were centred and scaled. Significance was evaluated at an alpha level of .05.

We chose to analyse the lexical knowledge variables as individual predictors instead of indicators of a factor in a structural equation model. This decision was based on previous

research (Zhang & Zhang, 2022) demonstrating that, although interrelated, lexical knowledge measures may predict comprehension differently.

It is common practice in L2 research to use beta coefficients and significant *p*-values as indices of predictor importance. Suppression effects, however, often arise in regression models (Hair et al., 2019), whereby the inclusion of one variable increases the predictive validity of another variable due to a higher degree of intercorrelatedness (but, critically, not multicollinearity) between predictor variables than between predictor variables and the outcome variable (Mizumoto, 2023). This suppressor effect can result in a disconnect between the magnitude of beta coefficients and the magnitude of the corresponding correlation coefficients.

Given the likelihood that the predictors would be highly intercorrelated and the impact that this would have on standardised beta coefficients, dominance analyses were conducted to estimate predictor importance accurately using the *calc.relimp()* function from the *relaimpo* package (Groemping & Matthias, 2021) for the linear models and the *domin()* function from the *dormir* package (Luchman, 2022) for the ordinal models. We interpreted non-overlapping confidence intervals around relative contributions as evidence of significant differences in importance between predictors. When confidence intervals overlapped, we considered the difference non-significant only if the confidence interval for the *difference* itself included zero (Grömping, 2006).

Results

Descriptive Statistics

Internal consistency reliability was generally good-to-excellent for the three knowledge measures ([Insert Table 1 approximately here]

and [Insert Table 2 approximately here]

). Descriptive statistics show that, on average, learners knew at least half of the words in each knowledge measure and that meaning-recognition was consistently the strongest, followed by form-recognition (once corrected for false alarm rates) and then form-recall, with non-overlapping CIs between any pair of measures. The range in accuracy across the knowledge measures indicated that our approach to word selection appropriately captured a range of words that learners knew to varying degrees. Henceforth, only the adjusted (not raw) form-recognition scores are used, as these accounted for guessing behaviour.

[Insert Table 1 approximately here]

[Insert Table 2 approximately here]

[Insert Table 3 approximately here]
[Insert Table 4 approximately here]

Assessing the construct validity of CV. Before examining relationships between knowledge, processing, and proficiency measures, we ascertained the extent to which CV measured the automaticity of word recognition. Spearman's *rho* (.365) indicated a small (bordering on medium) statistically significant, positive relationship between mean RT and CV. Thus, we refer to CV as a marker of processing stability *and* automaticity, henceforth.

[Insert Table 5 approximately here]

Relationships between knowledge and processing measures. Spearman's correlations (Table 5) revealed a high degree of interrelatedness ($r \geq .60$, Plonsky & Oswald, 2014)

between the knowledge measures (form-recall, meaning-recognition, form-recognition), ranging from .69 to .82, but not so strong to raise concerns about multicollinearity (r ≥. 90; Kline, 2023). Critically, however, we found little (if any) evidence of a relationship between the knowledge and processing measures: Where there was a significant correlation, the effect size was very small (*rho* ≤ . –.14). These initial analyses could provide preliminary evidence to suggest that at this early, fragile stage of proficiency, lexical knowledge and processing may be relatively independent. At the same time, the relatively small correlation may be indexing two opposing forces: That is, the learning of new words, which are initially processed with less stability, and the automatization of already established word representations. These two processes may effectively counteract to reduce the likelihood of finding a clearer association in one 'direction'.

Relationships between knowledge, processing, and proficiency measures. All three knowledge measures showed significant moderate-to-strong positive correlations with listening measures, ranging from .55 to .72, and reading measures, ranging from .57 to .75 (Table 5). However, there was no relationship between the processing and listening or reading measures.

Summary. We found a very high degree of interrelatedness both within the knowledge measures and between the knowledge and proficiency measures, regardless of the modality of comprehension. In contrast, there was very little (if any) evidence of any relationship between the knowledge and processing measures or between the processing and proficiency measures.

RQ1: Dimensionality of lexical mastery.

To investigate whether lexical knowledge and processing represented a unidimensional lexical mastery construct, we fitted a one-factor CFA model where the three knowledge

measures and the two processing measures loaded onto a single 'lexical mastery' factor. Methods effects resulting from the non-independence of mean RT and CV were accounted for by allowing the two processing measures to correlate.

The overall fit indices (Table 6) suggested that the one-factor model demonstrated a good overall fit: SRMR and CFI—our focus here, given the relatively small sample size (Kenny, 2015)—were both within the acceptable range (Brown, 2015; Hu & Bentler, 1999).

[Insert Table 6 approximately here]

[Insert Figure 3 approximately here]

Although the one-factor model demonstrated an acceptable fit with the data, it did not achieve convergent validity, as the Average Variance Extracted (.457) fell below the recommended threshold of .5 (Awang, 2012). The reliability of the one-factor model was also questionable. Although the composite reliability coefficient (.631) was slightly above the acceptable threshold (> .6 according to Fornell & Larcker, 1981), Cronbach's alpha (.527) fell below this threshold.

Inspection of the standardised regression coefficients (i.e., paths) between the lexical mastery factor and the two processing measures were weak, negative, and non-significant, suggesting further evidence of model misfit (Figure 3). This contrasted with the regression coefficients for the three knowledge measures, which were very high, positive, and significant ($p < .05$). The high loadings ($\beta = .81$ to $.92$) indicate a lack of discriminant validity between lexical mastery and the three knowledge types and suggest that the knowledge types represented a single underlying construct.

In sum, although the unidimensional model demonstrated an acceptable fit with the data, it did not reach convergent validity and was not found to be particularly reliable either.

Furthermore, inspection of standardised regression coefficients suggested that the lexical knowledge and processing measures did not measure the same construct. Instead, the lexical processing measures may have elicited a construct relatively independent of lexical knowledge or, at least, independent of the knowledge measures included in this study. In presenting these conclusions, we advocate caution, given that our findings are based only on an interpretation of a one-factor model as opposed to a comparison of a one-factor model with a two-factor one.

RQ2: Components of lexical mastery as predictors of listening and reading

Hierarchical regression models were computed to explore the extent to which types of lexical knowledge and processing contributed to L2 listening and reading, with dominance weights used to ascertain the relative importance of each predictor. Full model outputs are presented in Appendix S9.

Little evidence of multicollinearity was found between predictor variables (Variation Inflation Factor < 5 for each model). However, we applied Box-Cox transformations to the DELF reading data to meet the normality and homoscedasticity assumptions for the linear models. We also merged Levels 3^6 and below into one level to meet the proportional odds assumption for the ordinal models.

Listening models. Lexical knowledge as a global construct—combining form-recall, meaning-recognition, and form-recognition—explained 45.76% of the variance in DELF listening (see Step 1 in Table 13 of Appendix S9). Of these three predictors (see Table 14 of Appendix S9), form-recognition (19.49%, 95% CI [13.93%, 26.12%]) explained significantly more variance than form-recall (10.95%, 95% CI [7.11 %, 15.76%]). Form-recognition explained a similar amount of variance to meaning-recognition (15.32%, 95% CI [11.96%, 19.86%]). Although form-recall was not a significant predictor, its correlation with DELF

listening (*rho* = .550; see Table 5) and its predictor-specific R^2 (10.95 %, 95% CI [7.11 %, 15.76%]) suggested it explained some variance in DELF listening performance. The inclusion of CV or mean RT did not significantly increase the variance explained by the model (see model comparisons in Table 13 of Appendix S9; CV: F(1, 207) = 0.019, p = .890; RT: F(1, 207) = 0.021, p = .885).

A similar pattern emerged for the GCSE listening data. Lexical knowledge explained 53.20% of the variance (see Step 1 in Table 13 of Appendix S9), with form-recognition (22.49%) explaining the most, followed by meaning-recognition (15.59%) and then form-recall (15.12%), even though form-recognition and form-recall were significant in the model and meaning-recognition was not (see Table 20 of Appendix S9). Neither CV nor mean RT significantly improved model fit (see model comparisons in Table 13 of Appendix S9; CV: $\chi^2(1) = 0.338$, p = .561; RT: $\chi^2(1) = 0.720$, p = .396).

Reading models. Lexical knowledge explained 54.80% of the variance in DELF reading, slightly higher than in listening (see Step 1 in Table 13 of Appendix S9). Reflecting a similar order to listening, form-recognition (22.77%, 95% CI [17.35%, 28.66%]) accounted for the most variance, closely followed by meaning-recognition (20.35%, 95% CI [15.85%, 25.31%]), and then form-recall (11.69%, 95% CI [7.33%, 17.31%]). Both form-recognition and meaning-recognition explained significantly more variance than form-recall. (Note that form-recall did not reach significance, see Table 17 of Appendix S9.) As with the listening models, the inclusion of CV or RT did not significantly improve the model fit (see model comparisons in Table 13 of Appendix S9; CV: F(1, 211) = 0.525, p = .470; RT: F(1, 211) = 0.165, p = .686).

A similar pattern was found for the GCSE reading (see Table 23 of Appendix S9), with lexical knowledge accounting for 57.00% of the variance. Form-recognition (24.54%) explained the most variance, followed by meaning-recognition (19.90%) and then form-recall

(12.57%). As with the DELF listening and reading models, form-recall did not reach significance. However, alternative metrics of predictor importance that are not as sensitive to suppression effects, including Spearman's correlations (*rho* = .584) and the predictor-specific R^2 (12.57%), suggested a high degree of interrelatedness between form-recall and reading comprehension. Again, CV or RT did not improve the model fit (see model comparisons in Table 13 of Appendix S9; CV: $\chi^2(1) = 0.207$, $p = .886$; RT: $\chi^2(1) = 0.735$, $p = .391$).

Summary. Our findings broadly suggest that lexical knowledge—both as a construct and its components—played a critical role in predicting listening and reading when assessed via low-stakes (DELF) *and* high-stakes (GCSE) tests. On the other hand, processing measures did not explain any unique variance in comprehension.

Discussion

In this study, we investigated the extent to which types of lexical knowledge and processing (a) are relatively independent and (b) predict listening and reading among 218 beginner-to-low-intermediate English-speaking classroom learners of French. We now discuss key findings and potential implications for the conceptualisation, learning, teaching, and assessment of vocabulary.

RQ1: Dimensionality of lexical mastery: Lexical knowledge and processing

Our findings aligned with those from previous studies (e.g., González-Fernández, 2022; González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004; Webb, 2005) concerning (a) the high degree of interrelatedness between types of lexical knowledge and (b) the hierarchy of difficulty between types of form–meaning knowledge: meaning-recognition scores were consistently the highest, followed by form-recognition (once corrected for false alarm rates)[7], and then form-recall. These findings pertained with a larger number of items (*k*

= 50) relative to previous research, increasing the generalisability of the findings. Additionally, our items were selected to enhance learner-relevance: All words appeared in the lists associated with the national GCSE French curricula, thus adding to the ecological validity of our findings. Given this tightly controlled set of items, the observed differences between knowledge types are arguably particularly striking. Whereas previous studies reported these effects using fewer items from a broader range of frequencies (e.g., González-Fernández & Schmitt, 2020), we observed a similar hierarchy in performance even when sampling from a restricted frequency band (0-2,000) and only from items with which we expected most participants to be familiar to some extent.

Critically, form-recognition (i.e., lexical-decision accuracy corrected for false alarms) did not correlate with either mean RT (speed) or CV (stability and automaticity) for correct trials on the same word recognition task. Although form-recall correlated with mean RT, the effect size was small ($r = -.14$), and no correlations were found with CV. These findings are noteworthy for two reasons. First, the lexical-decision—yielding measures of speed, stability, and automaticity of word recognition—was administered approximately 24 hours after the form-recall and meaning-recognition task. As such, priming effects may have occurred, potentially improving RTs and/or automaticity and strengthening any possible relationship between lexical knowledge and processing. Nevertheless, despite the potential for priming, lexical processing did *not* appear to correlate with lexical knowledge, at least among the learners tested in this study.

Second, our findings differ quite substantially from previous studies (e.g., Harrington, 2018; Pellicer-Sánchez & Schmitt, 2012) that found a significant relationship between lexical knowledge and word recognition speed for correct trials. As such, they do not align with Harrington's (2018) Lexical Facility proposal that knowledge and processing are complementary, interrelated indices of L2 lexical mastery. One potential explanation for

these different observations could be that our test items were too easy, producing a ceiling effect and thus rendering insufficient variance. The only other study, to the best of our knowledge and according to Harrington (2018), to not observe a correlation between knowledge and mean RT was Miralpeix and Meara (2010), who tested intermediate and advanced L2 English students, using a separate vocabulary size test and lexical-decision task. Harrington suggested that the lack of correlation observed in Miralpeix and Meara's study was due to the ease of the target words, which also seemed highly automatised in participants' mental lexicons, perhaps leaving no variance for correlations: The mean RT for their intermediate and advanced L2 learners was 815ms, only 31ms slower than their native speakers. We suspect this reason is unlikely to explain our findings: We compared the mean RTs for 48 of the 50 words in our study with the open dataset from the French Lexicon Project (Ferrand et al., 2010) and found that the mean RTs of our beginner-to-low-intermediate learners were much slower than those of native speakers (823ms against 633ms). Along with the mean 65% form–recognition (adjusted) accuracy rate, this does not indicate a strong ceiling effect that could have masked a potential correlation.

Our findings instead suggest that at this early stage of language learning, lexical knowledge and processing could be relatively independent, rather than interrelated, dimensions of lexical mastery. The Lexical Facility proposal may be more applicable at intermediate and advanced stages when representations of known words are more readily and reliably activated, thus producing significant associations between knowledge and speed, stability, and automaticity. Indeed, evidence in support of the Lexical Facility proposal comes from intermediate and advanced learners (i.e., Harrington, 2018; Pellicer-Sánchez & Schmitt, 2012). This is perhaps due to the methodological risks of testing beginners, as low accuracy rates can lead to a loss in RT data. However, as in our study, these risks can be mitigated by selecting words that beginners might be expected to have at least some knowledge of.

A more accurate explanation for our findings may be Perfetti's (2007) Lexical Quality Hypothesis, whereby processing is a by-product of complete representations (i.e., knowledge). As such, processing could represent a related dimension of lexical mastery when lexical representations are more consolidated and reliable than those of our participants. Perhaps our participants' representations of the words were insufficiently established, as few learners performed at ceiling on the knowledge measures (form-recall, form-recognition, and meaning-recognition). Moreover, mean RT only weakly correlated (.37) with the CV, in line with Hui and Godfroid (2021) who reported a correlation coefficient of .35 in their lexical task, potentially suggesting little evidence of automaticity within the sample of learners tested. In other words, lexical knowledge may not be fully automatised at this proficiency level, weakening relations between knowledge and processing measures.

An important, complementary explanation for the lack of relationship between lexical knowledge and processing among our learners, as suggested by an anonymous reviewer, might relate to two opposing forces—knowledge accumulation and automatization—having an apparent neutralising effect on CV. This neutralising effect may also explain why CV predicted very little variance in listening and reading proficiency scores. To elaborate, our learners were still at the very early stages of acquisition. Many new words were still being learned (accumulated), and knowledge of existing words was still being consolidated (automatised). Indeed, Hulstijn et al. (2009, p. 555) state that "knowledge accumulation forms part of skill acquisition because, in real L2 learning, exposure to new words goes hand in hand with exposure to words encountered previously". Given learners' low amount of instruction to date, a higher accuracy score on the form-recognition measure (which provided the basis of the CV calculations, given the inclusion of correct trials only) suggests that a learner may have relatively *recently* acquired more words in their lexicons than a learner with a lower accuracy score.

Lexical representations for these relatively recently acquired "additional" words are likely to have been unstable, causing CV to be higher. Indeed, CV has been found to increase initially in one's developmental trajectory, contrary to an expectation that it would always go down as learning progressed (e.g., Hui, 2020; Solovyeva & DeKeyser, 2018). At the same time, learners' representations of other (perhaps less recently acquired) words were likely to be comparatively more stable in their lexicons and possibly even automatised, causing CV to be lower for these words. However, any evidence of automaticity (which would have rendered a lower CV and stronger correlation between mean RT and CV) may have been masked by knowledge accumulation having a neutralising effect on any lower CVs that *could* have resulted from automatization and consequently *could* have predicted listening and reading proficiency scores.

In sum, our findings suggest that at this early stage of proficiency, lexical knowledge and processing did not load onto the same construct and behaved more like relatively independent dimensions of lexical mastery. However, further research is needed to investigate the extent to which any possible distinction between lexical knowledge and processing varies as a function of proficiency and/or perhaps knowledge and processing of individual words, whereby some words render a strong relation, and others do not.

RQ2: Predictive validity of lexical knowledge and processing measures of lexical mastery for listening and reading comprehension

Our second aim was to investigate the extent to which knowledge and processing measures predict L2 listening and reading. A consistent finding that replicated across low-stakes (DELF) and high-stakes (GCSE) proficiency tests was that knowledge measures, both individually and combined, predicted listening and reading. However, findings for form-

recall were more nuanced (see below). Our findings thus align generally with previous studies, including Zhang and Zhang's (2022) extensive meta-analysis.

These findings also underscore the importance of high-frequency vocabulary: We observed that knowledge—albeit partial in some cases—of a relatively small set of words (k = 50) predicted between 46% and 53% and 55% and 57% of the variance in listening and reading, respectively. Teaching high-frequency words, at least at these proficiency levels in low-exposure, instructed contexts, may, therefore, support listening and reading development.

We further found that form-recognition explained the most variance in listening *and* reading, followed by meaning-recognition and then form-recall. However, CIs around the variance in listening and reading explained by each type generally overlapped, suggesting that differences in the predictive importance of each type were relatively subtle. As such, we do not draw strong implications for any potential order of importance for teaching or testing. What was clear, however, was the usefulness of drawing on multiple metrics to assess predictor importance, as Mizumoto (2023) recommended. For instance, form-recall was not a significant predictor in three of the four models. Yet, its dominance weights (i.e., predictor-specific R^2 values), albeit small in magnitude relative to other predictors, and Spearman's *rho* often overlapped in CIs with predictors that did reach significance in the models.

We note that the observed descriptive patterns do not clearly align with previous research that found (a) for listening, form-recall was the *strongest* predictor, followed by meaning-recall (not tested in our main study for practical reasons) and then meaning-recognition, and (b) for reading, meaning-recall (not tested in our main study) was the strongest predictor, followed by form-recall and then meaning-recognition (Zhang & Zhang, 2022).

However, the strong predictive role of meaning-recall for both listening and reading was reported in our sub-study (for more details, see Appendix S10). In this sub-study, we

tested 87 learners (from the main study) on their meaning–recall knowledge of 60 high-frequency words, including 30 of the critical words from the main study. These analyses revealed that the inclusion of meaning-recall significantly improved the variance explained in the models of DELF reading (with R^2 increasing from 54.1% to 60.5%), GCSE listening (with R^2 increasing from 53.6% to 60.9%), and GCSE reading (with R^2 increasing from 51.1% to 55.4%).

The importance of meaning–recall relative to other types of lexical knowledge for reading comprehension broadly mirrored those reported in previous studies (Zhang & Zhang, 2022). That is, of the four knowledge measures, meaning–recall explained the most variance (20.65%) in DELF reading, followed by form-recognition (17.81%), meaning-recognition (14.39%), and then form-recall (7.52%). Similarly, meaning-recall and meaning-recognition explained similar amounts of variation (16.84% and 16.18%, respectively) in GCSE reading, followed by form-recognition (14.87%) and form-recall (7.52%).

In contrast to previous studies, where form-recall had the strongest correlation with L2 listening comprehension, form-recognition explained the most variance (19.38%) in DELF listening, followed by meaning-recognition (15.17%), meaning-recall (14.48%), and then form-recall (6.03%). Furthermore, meaning–recall explained the most variance (22.96%) in GCSE listening, followed by form-recognition (14.01%), meaning-recognition (13.27%), and then form-recall (8.08%)

There are several possible reasons why our observed order of predictive importance (at least, as reported in the main study) did not align with the general trends meta-analysed by Zhang and Zhang (2022). First, some of the trends observed by Zhang and Zhang were nuanced, with overlapping CIs suggesting potential variability between studies. Second, form-recall was the weakest knowledge type among our learners. Previous studies have also found that form-recall tends to elicit the lowest scores (e.g., González-Fernández, 2022;

González-Fernández & Schmitt, 2020; Laufer & Goldstein, 2004; Webb, 2005). However, it may be that knowledge (of any type) must exceed a certain threshold before it is robust enough to play a reliable or strong role in comprehension. However, this idea is speculative, as little research—none, to our knowledge—can inform us about the early stages of language learning when knowledge is fragile.

Third, the predictive role of each knowledge type may vary as a function of age and/or proficiency. Zhang and Zhang (2022) reported that lexical knowledge predicted reading most strongly among university students (r = .61, 95% CI [.54, .67]), followed by secondary school students (r = .59, 95% CI [.51, .67]), and then elementary school students (r = .53, 95% CI [.48, .58]). Interestingly, a reverse pattern was found for listening, with lexical knowledge predicting performance most strongly among elementary school students (r = .60, 95% CI [.43, .62], followed by secondary school students (r = .54, 95% CI [.40, .66]), and then university students (r = .53, 95% CI [.43, .62]). However, wide, highly overlapping CIs indicate a range of findings between individual studies and nuanced differences between ages (and, by proxy, education levels). Furthermore, these findings related to *overall* lexical knowledge rather than individual types. Further research is therefore needed to ascertain the extent to which distinct types of lexical knowledge differentially predict listening and reading.

Third, unlike previous studies that sampled from a broader range of frequency bands, our critical items were sampled from a bank of high-frequency words specific to the participants' curriculum. It may be that the effect of lexical knowledge type on comprehension varies as a function of word frequency. For example, recognition of high-frequency words may contribute more strongly to L2 listening and reading than form-recall knowledge of the same words. As far as we know, however, no research has investigated this question, and further research is needed.

Regarding the contribution of processing measures, we found no evidence that speed, stability, and automaticity of word recognition—at least of the words tested in this study—had a relationship with listening and reading. Our findings align with several others indicating that speed, stability, and automaticity of word recognition have little—if any—relationship with L2 reading (Alshehri & Zhang, 2022; van Gelderen et al., 2004) or listening (Andringa et al., 2012), at least relative to lexical knowledge.

One potential explanation could relate to the idea that fluent lexical processing can only be achieved once representations are (more) complete and/or once the (size of the) lexicon has become more complete and/or fully established, in line with Perfetti's (2007) Lexical Quality Hypothesis. That is, lexical representations must reach a certain threshold (e.g., of reliability, association strength, or activation levels), and the lexicon must reach a specific size before automatic recognition can be achieved and thus influence actual listening and reading comprehension. However, given the paucity of studies investigating the interconnectedness of lexical knowledge and processing and listening and reading proficiency, further research is needed to explore such claims.

Limitations

Our battery of measures was not fully comprehensive. We wanted to avoid fatigue and attrition within and between sessions, with a hard-to-reach population giving up their vacation time immediately after their high-stakes national exams. Two obvious gaps are that we did not test meaning-recall—at least not in the main study—or aural lexical knowledge and processing. In addition to our practical constraints, testing all subcomponents in both modalities on the same set of words would have been methodologically controversial. If possible, though, it would have provided insight into whether relations between knowledge, processing, and comprehension vary as a function of the modality in which they are measured. Furthermore, the role of modality may depend on (a) phoneme-grapheme

correspondence transparency in the L2, (b) cross-linguistic differences in this regard, and (c) individuals' knowledge of phoneme-grapheme correspondences, all of which may influence the extent to which lexical knowledge in the written modality can serve comprehension in the aural modality, and vice versa. We encourage further research to explore these avenues across different proficiencies and L1–L2 pairings.

An additional limitation is that we used a single time-sensitive measure of lexical processing (i.e., a lexical-decision) and that this measure only tapped into form-recognition—not meaning-recognition. However, we found (a) no relation between knowledge and processing measures from the same task (form-recognition) and (b) strong correlations between form-recognition and the other two knowledge measures. Therefore, we suspect that, for these learners, if we had taken processing measures for a different knowledge type (e.g., meaning-recognition), we would have also found no observable relations between processing and knowledge. Nevertheless, we concur with others, including Godfroid (2020), that L2 vocabulary studies could usefully integrate different time-sensitive measures (such as eye-tracking but also timed matching or semantic judgment tasks) despite the challenges especially among hard-to-reach populations and for non-laboratory settings.

Conclusion

This study investigated the extent to which knowledge and processing measures of lexical mastery (a) are relatively independent and (b) predict (both individually and combined) L2 listening and reading among 218 beginner-to-low-intermediate English-speaking classroom learners of French. Our findings indicated that the three knowledge measures (form-recall, meaning-recognition, form-recognition) were highly interrelated but relatively independent from processing measures (speed, stability, and automaticity of word recognition). All three knowledge measures—together and individually—predicted listening and reading, though findings for form-recall were more nuanced, whereas processing measures were not

associated with comprehension scores. We argued that this separate conceptualisation of knowledge and processing, their (lack of) relations with L2 receptive proficiency, and relatively weak correlations with mean RT and CV are likely to be artefacts of learners' partial lexical representations and growing lexicons and, thus, provide limited evidence of automatization of lexical knowledge. We suggested that the dimensionality of lexical mastery may vary as a function of proficiency, with lexical mastery becoming more unidimensional—i.e., lexical knowledge and processing potentially relating more closely to each other—as learners become more proficient. Future research is needed, though, to validate these claims with learners from different proficiency levels, L1-L2 combinations, and processing measures.

Endnotes

[1] In this article, we use "lexical mastery" as an umbrella term—frequently used by Laufer (2021) and Schmitt (2019)—to refer to mastery of lexical knowledge and processing.

[2] Curcin and Black (2019) estimate that a Level 4 (a pass) or above in GCSE French is equivalent to A1/A2 on the Common European Framework of Reference for Languages' six-point scale.

[3] Analyses presented in Appendix S6 show that year of data collection had no effect on performance.

[4] Measures were put in place to discourage participants from cheating. These measures included disabling any copy and paste functions, forcing web browsers into full screen mode, and a warning at the beginning of each task that if participants consulted external sources, they would not receive the voucher for participation.

[5]These global indices have been shown to be less sensitive to sample size than the root mean square error of approximation (RMSEA; Kenny, 2015).

[6] Level 4 or above is considered a pass at GCSE.

[7]Although raw form-recognition accuracy was high (M = 89.86%, 95% CI [88.64%, 91.08%]), this was most likely due to positively skewed guessing, driven by the orthographic overlap with real words: Once corrected for false alarms, accuracy significantly decreased (M = 64.97%, 95% CI [62.38%, 67.56%]).

References

Alshehri, M.G., & Zhang, D. (2022). The lexical basis of second language reading comprehension: From (sub)lexical knowledge to processing efficiency. *Language Learning*, *72*(2), 325-364. https://doi.org/10.1111/lang.12478

ALTE. (2023). *ALTE framework 2023*. https://cdn.wildapricot.com/218795/resources/Documents/minimum_standards_en.pdf?version=1495011131000&Policy=eyJTdGF0ZW1lbnQiOiBbeyJSZXNvdXJjZSI6Imh0dHBzOi8vY2RuLndpbGRhcHJpY290LmNvbS8yMTg3OTUvcmVzb3VyY2VzL0RvY3VtZW50cy9taW5pbXVtX3N0YW5kYXJkc19lbi5wZGY~dmVyc2lvbj0xNDk1MDExMTMxMDAwIiwiQ29uZGl0aW9uIjp7IkRhdGVMZXNzVGhhbiI6eyJBV1M6RXBvY2hUaW1lIjoxNzQyMDQyNjk5fSwiSXBBZGRyZXNzIjp7IkFXUzpTb3VyY2VJcCI6IjAuMC4wLjAvMCJ9fX1dfQ__&Signature=ngY4huNvGjHUh-7dZlYsmtWDVxyj~Y~VZk8~Zq1vcoyznE1V~GuhumWedpGgmhEBX8Qu-s57zh0ZIO-f3jfMPnV~r5mWjiTlVG29BkdzGKWEZIrHl6CXazyYY5XMfuF07Z~PqAItXz363LrZdlOHolHohmhMdSpb7WJ0dQrZdiP5Y5CpwqaMJR6Q~Q~bn3ltOHqaMVFlyTmbSzRiwMeWFs35IlkCrQ6EGfFplP8D9Y5x4GUA3TcgRG-TypXuGUfUwZjPQ4GbQGe5yVGEDfl7v~FvRK4zfxKm6sZ7g-

ue6TUbr3NwmRukNmjMpOzSxu5oDvfUW8LKKduS~z7AWrl9uA__&Key-Pair-Id=K27MGQSHTHAGGF

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, *62*(s2), 49-78. https://doi.org/10.1111/j.1467-9922.2012.00706.x

AQA. (2016). *GCSE French (8658) specification*. https://filestore.aqa.org.uk/resources/french/specifications/AQA-8658-SP-2016.PDF

Awang, Z. (2012). *A handbook on SEM structural equation modelling: SEM using AMOS graphic* (5th edition). Universiti Teknologi Mara Kelantan.

Batista, R., & Horst, M. (2016). A new receptive Vocabulary Size Test for French. *The Canadian Modern Language Review*, *72*(2), 211-233. https://doi.org/10.3138/cmlr.2820

Bordag, D., Gor, K., & Opitz, A. (2022). Ontogenesis model of the L2 lexical representation. *Bilingualism: Language and Cognition*, *25*(2), 185-201. https://doi.org/10.1017/S1366728921000250

Brown, T.A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Carrillo, M. S., Alegría, J., & Marín, J. (2013). On the acquisition of some basic word spelling mechanisms in a deep (French) and a shallow (Spanish) system. *Reading and Writing*, *26*(6), 799-819. https://doi.org/10.1007/s11145-012-9391-6

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, *31*(5), 693-713. https://doi.org/10.1093/applin/amq031

Curcin, M., & Black, B. (2019). Investigating standards in GCSE French, German and Spanish through the lens of the CEFR. In *Ofqual*.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/844034/Investigating_standards_in_GCSE_French__German_and_Spanish_through_the_lens_of_the_CEFR.pdf

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2022). Formal versus informal L2 learning. *Studies in Second Language Acquisition*, *44*(1), 87-111. https://doi.org/10.1017/S0272263121000097

Dudley, A., Marsden, E., & Bovolenta, G. (2024). A Context-Aligned Two Thousand Test: Toward estimating high-frequency French vocabulary knowledge for beginner-to-low intermediate proficiency adolescent learners in England. *Language Testing*, *41*(4), 759-791. https://doi.org/10.1177/02655322241261415

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488-496. https://doi.org/10.3758/BRM.42.2.488

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50. https://doi.org/10.1177/002224378101800104

France Éducation International. (n.d.). *DELF junior/scolaire*. Retrieved December 21, 2022, from https://www.france-education-international.fr/en/diplome/delf-junior-scolaire?langue=en

Godfroid, A. (2020). Sensitive measures of vocabulary knowledge and processing. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 433-453). Routledge.

González-Fernández, B. (2022). Conceptualizing L2 vocabulary knowledge: An empirical examination of the dimensionality of word knowledge. *Studies in Second Language Acquisition*, 1-31. https://doi.org/10.1017/S0272263121000930

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*(4), 481-505. https://doi.org/10.1093/applin/amy057

Groemping, U., & Matthias, L. (2021). *relaimpo: Relative importance of regressors in linear models* (Version 2.2-6). https://cran.r-project.org/web/packages/relaimpo/index.html

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17*(1). https://doi.org/10.18637/jss.v017.i01

Gyllstad, H., McLean, S., & Stewart, J. (2021). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*, *38*(4), 558-579. https://doi.org/10.1177/0265532220979562

Hair, J.F., Babin, B.J., Anderson, R.E., & Black, W.C. (2019). *Multivariation data analysis* (8th edition). Cengage.

Harrington, M. (2018). *Lexical facility: Size, recognition speed and consistency as dimensions of second language vocabulary knowledge*. Palgrave Macmillan.

Hawkes, R., & Lillington, C. (2016). *Viva! Edexcel GCSE (9-1) Spanish Higher Student Book*. Pearson Education.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hui, B. (2020). Processing variability in intentional and incidental word learning: An extension of Solovyeva and DeKeyser (2018). *Studies in Second Language Acquisition*, *42*(2), 327-357. https://doi.org/10.1017/S0272263119000603

Hui, B., & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. *Applied Psycholinguistics*, *42*(5), 1089-1115. https://doi.org/10.1017/S0142716420000193

Hui, B., Godfroid, A., & Elgort, I. (2025). A construct validation study of time-sensitive word-knowledge measures, *Applied Linguistics*. https://doi.org/10.1093/applin/amaf037

Hui, B., & Jia, R. (2024). Reflecting on the use of response times to index linguistic knowledge in SLA. *Annual Review of Applied Linguistics*, 44, 45-55. https://doi.org/10.1017/S0267190524000047

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*(3), 227-245. https://doi.org/10.1191/0265532202lt229oa

Hulstijn, J.H., van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*(4), 555-582. https://doi.org/10.1017/S0142716409990014

In'nami, Y., Koizumi, R., Jeon, E.H., & Arai, Y. (2022). Chapter 8. L2 listening and its correlates. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigators* (pp. 235-283). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.13.08inn

Jeon, E. H., & Yamashita, J. (2022). Chapter 3. L2 reading comprehension and its correlates. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp. 29-86). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.13.03jeo

Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122-149. https://doi.org/10.1037/0033-295X.99.1.122

Kenny, D. A. (2015). *Measuring model fit*. http://davidakenny.net/cm/fit.htm

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th edition). Guilford Press.

Laufer, B. (2021). Lexical thresholds and alleged threats to validity: A storm in a teacup? *Reading in a Foreign Language*, *33*(2), 238-246. https://eric.ed.gov/?id=EJ1317203

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399-436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Paribakht, T.S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*(3), 365-391. https://doi.org/10.1111/0023-8333.00046

Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners*. Routledge. https://doi.org/10.4324/9780203883044

Luchman, J. (2022). *domir: Tools to support relative importance analysis* (Version 1.0.0). https://cran.r-project.org/web/packages/domir/index.html

Marsden, E., Dudley, A., & Hawkes, R. (2023). Use of word lists in a high-stakes, low-exposure context: Topic-driven or frequency-informed. *Modern Language Journal*, 107(*3*), 669–692. https://doi.org/10.1111/modl.12866

McLean, S., Matthews, J., & Milliner, B. (2024). Listening and lexical knowledge. In E. Wagner, A. O. Batty, & E. Galaczi (Eds.), *The Routledge handbook of second language acquisition and listening* (pp. 146-160). Routledge. https://doi.org/10.4324/9781003219552-13

McManus, K., & Marsden, E. (2019). Signatures of automaticity during practice: Explicit instruction about L1 processing routines can improve L2 grammatical processing. *Applied Psycholinguistics*, *40*(1), 205-234. https://doi.org/10.1017/S0142716418000553

Miralpeix, I., & Meara, P. (2010). *The written word*. www.lognostics.co.uk/Vlibrary

Mizumoto, A. (2023). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning*, *73*(1), 161-196. https://doi.org/10.1111/lang.12518

Nation, P. (2013). *Learning vocabulary in another language*. Cambridge University Press.

Nation, P. (2020). The different aspects of vocabulary knowledge. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 15-29). Routledge.

Pearson Edexcel. (2018). *GCSE French (1FR0) specification*. https://qualifications.pearson.com/content/dam/pdf/GCSE/French/2016/specification-and-sample-assessments/Specification-Pearson-Edexcel-Level-1-Level-2-GCSE-9-1-French.pdf

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes-No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*(4), 489-509. https://doi.org/10.1177/0265532212438053

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357-383. https://doi.org/10.1080/10888430701530730

Plonsky, L., & Oswald, F.L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912. https://doi.org/10.1111/lang.12079

R Development Core Team. (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. http://www.rproject.org

Ripley, B., Venables, B., Bates, D.M., Hornik, K., & Gebhardt, A. (2022). *MASS: Support functions and datasets for venables and Ripley's MASS* (7.3-58.1). https://cran.r-project.org/web/packages/MASS/index.html

Rosseel, Y. (2023). *lavaan: Latent variable analysis* [R package]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/package=lavaan

Saito, K., Uchihara, T., Takizawa, K., & Suzukida, Y. (2023). Individual differences in L2 listening proficiency revisited: Roles of form, meaning, and use aspects of phonological vocabulary knowledge. *Studies in Second Language Acquisition*, *47*(1), 26-52. https://doi.org/10.1017/S027226312300044X

Schmalz, X., Marinus, E., Coltheart, M., & Castles, A. (2015). Getting to the bottom of orthographic depth. *Psychonomic Bulletin & Review*, *22*(6), 1614-1629. https://doi.org/10.3758/s13423-015-0835-2

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261-274. https://doi.org/10.1017/S0261444819000053

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Segalowitz, S.J., Segalowitz, N.S., & Wood, A.G. (1998). Assessing the development of automaticity in second language word recognition. *Applied Psycholinguistics*, *19*(1), 53-67. https://doi.org/10.1017/S0142716400010572

Solovyeva, K., & DeKeyser, R. (2018). Response time variability signatures of novel word learning. *Studies in Second Language Acquisition*, *40*(1), 225-239. https://doi.org/10.1017/S0272263117000043

Stanovich K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. Guilford Press.

Stewart, J., McLean, S., & Batty, A. (2021). Correlations of modalities of written vocabulary knowledge to listening and reading proficiency: A comparison. *Vocabulary Learning and Instruction*, *10*(2), 55-63. https://doi.org/10.7820/vli.v10.2.Stewart

van Gelderen, A., Schoonen, R., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential

analysis. *Journal of Educational Psychology*, *96*(1), 19-30. https://doi.org/10.1037/0022-0663.96.1.19

Vandenberghe, B., Perez, M.M., Reynvoet, B., & Desmet, P. (2021). Combining explicit and sensitive indices for measuring L2 vocabulary learning through contextualized input and word-focused instruction. *Studies in Second Language Acquisition*, *43*(5), 1009-1039. https://doi.org/10.1017/S0272263120000431

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33-52. https://doi.org/10.1017/S0272263105050023

Yanagisawa, A., & Webb, S. (2019). Measuring depth of vocabulary knowledge. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 371-386). Routledge.

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*(4), 696-725. https://doi.org/10.1177/1362168820913998

Zhang, X., Liu, J., & Ai, H. (2020). Pseudowords and guessing in the Yes/No format vocabulary test. *Language Testing*, *37*(1), 6-30. https://doi.org/10.1177/0265532219862265

Supporting Information

Appendix S1. Nation's framework of word knowledge and Godfroid's expansion

Appendix S2. Information about the GCSE in England

Appendix S3. Critical words

Appendix S4. Task instructions

Appendix S5. Form recall mark scheme

Appendix S6. Effect of year of data collection on lexical knowledge and processing

Appendix S7. Spearman's correlations with 95% confidence intervals

Appendix S8. Confirmatory factor analyses

Appendix S9. Hierarchical linear (for DELF) and ordinal (for GCSE) regression models

Appendix S10. Meaning recall analyses

Appendix S11. References

Tables and Figures

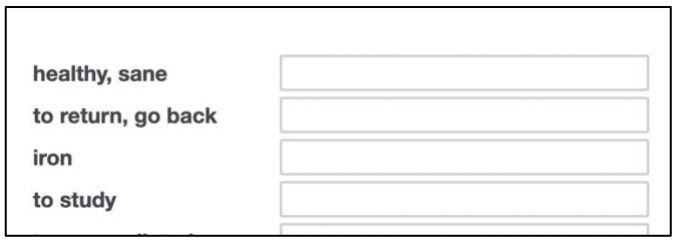

Figure 1. Sample items from the form-recall task.

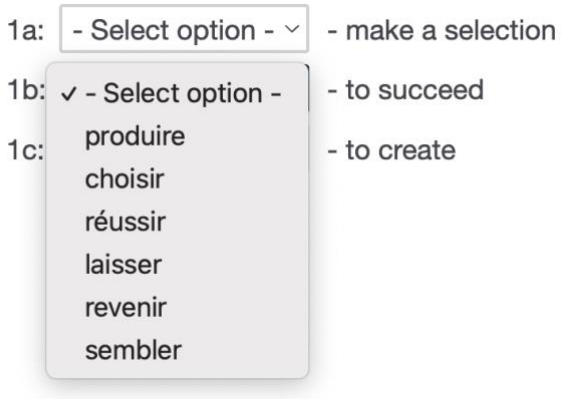

Figure 2. Sample cluster from the Two Thousand Test.

Table 1. Descriptive statistics and internal consistency reliability for the knowledge measures.

	Form Recall	Meaning-recognition	Form-recognition (Raw)	Form-recognition (Adjusted[1])
n	218	218	218	218
Mean (%)	58.75%	85.63%	89.86%	64.97%
SD (%)	19.68%	13.75%	9.14%	19.39%
Median (%)	60.50%	90.00%	92.00%	67.58%
95% CI (%)	[56.13%, 61.38%]	[83.80%, 87.47%]	[88.64%, 91.08%]	[62.38%, 67.56%]
Min (%)	6.00%	28.00%	54.55%	16.32%
Max (%)	98.00%	100.00%	100.00%	100.00%
Skew	-0.34	-1.63	-1.37	-0.50
Kurtosis	-0.34	3.10	1.87	-0.56
Omega	.94	.93	.83	-
Alpha	.93	.91	.82	-

[1] I_{SDT} = index of signal detection.

Table 2. Descriptive statistics and internal consistency reliability for the processing measures.

	Mean Response Time (in ms)	Coefficient of Variation (CV)
n	218	218
Mean	824.32	0.33
SD	132.58	0.08
Median	814.16	0.34
95% CI	[806.62, 842.02]	[0.32, 0.35]
Min	598.06	0.17
Max	1,764.28	0.56
Skew	1.80	0.12
Kurtosis	10.05	-0.59
Split Half	.80 [.76, .83]	-
Spearman Brown	.89 [.86, .91]	-

Note: Calculations for mean RT and CV were based only on response times for Hits (correct yes responses on the form-recognition test).

Table 3. Descriptive statistics and internal consistency reliability for the DELF proficiency measures.

	Listening	Reading
n	212	216
Mean	53.57%	71.69%
Median	52.00%	76.00%
SD	23.14%	22.99%
95% CI	[50.43%, 56.70%]	[68.60%, 74.77%]
Min	12.00%	12.00%
Max	100.00%	100.00%
Skew	0.18	-0.62
Kurtosis	-1.07	-0.72
Omega	.85	.90
Alpha	.85	.90

Note: Data from participants who achieved scores below 10% (n = 8 for the listening test; n = 4 for the reading test) were excluded from the dataset, as such low performance may indicate lack of engagement in the task.

Table 4. Descriptive statistics for French GCSE level.

	Percentage of learners achieving each level								
	U	3	4	5	6	7	8	9	Total[1]
Reading	2%	4%	4%	14%	5%	8%	19%	44%	194
Listening	3%	2%	6%	17%	4%	21%	20%	28%	194

[1] 89% (194) of the 218 participants sent an individual breakdown of their GCSE results.

Table 5. Spearman's correlational analyses between the knowledge, processing, and proficiency measures (with *p* values in brackets).

	Form-recall	Meaning-recognition	Form-recognition (raw)	Form-recognition (adjusted)	Mean response time	Coefficient of variation	DELF list.	DELF read.	GCSE list.
Meaning-recognition	**.749 (<.001)**								
Form-recognition (raw)	**.687 (<.001)**	**.753 (<.001)**							
Form-recognition (adjusted)	**.724 (<.001)**	**.822 (<.001)**	**.911 (<.001)**						
Mean response time	**-.139 (.040)**	-.131 (.053)	**-.137 (.044)**	-.100 (.141)					
Coefficient of variation	-.071 (.299)	-.105 (.122)	-.032 (.640)	-.077 (.255)	**.365 (<.001)**				
DELF listening	**.550 (<.001)**	**.688 (<.001)**	**.645 (<.001)**	**.666 (<.001)**	-.065 (.350)	-.087 (.205)			
DELF reading	**.573 (<.001)**	**.711 (<.001)**	**.652 (<.001)**	**.704 (<.001)**	-.066 (.331)	-.118 (.085)	**.717 (<.001)**		
GCSE listening	**.622 (<.001)**	**.705 (<.001)**	**.698 (<.001)**	**.718 (<.001)**	-.118 (.103)	-.077 (.290)	**.733 (<.001)**	**.697 (<.001)**	
GCSE reading	**.584 (<.001)**	**.709 (<.001)**	**.679 (<.001)**	**.752 (<.001)**	-.104 (.149)	-.099 (.172)	**.633 (<.001)**	**.664 (<.001)**	**.791 (<.001)**

Confidence intervals for each correlation are presented in Table 11 of Appendix S7. Small: $r > .25$; medium: $r > .40$; large: $r > .60$ (Plonsky & Oswald, 2014).

Table 6. Fit indices for the confirmatory factor analyses.

	$\chi2$	df	*p* value	$\chi2$ / df	CFI	RMSEA [90% CI]	SRMR	AIC	BIC	Adjusted BIC
Acceptable fit[1]			> .05	1-3	> .95	< .05	< .08	The smaller, the better		
One-factor	6.703	4	.152	1.676	.994	.053 [.000, .120]	.021	2,680.133	2,717.363	2,682.504

[1] As per guidelines specified by Hui and Bentler (1999) and Brown (2015). ***Note**:* CFI (Comparative fit index), RMSEA (root mean square error of approximation), SRMR (standardised root mean squared residual), AIC (Akaike information criterion), and BIC (Bayesian information criterion).

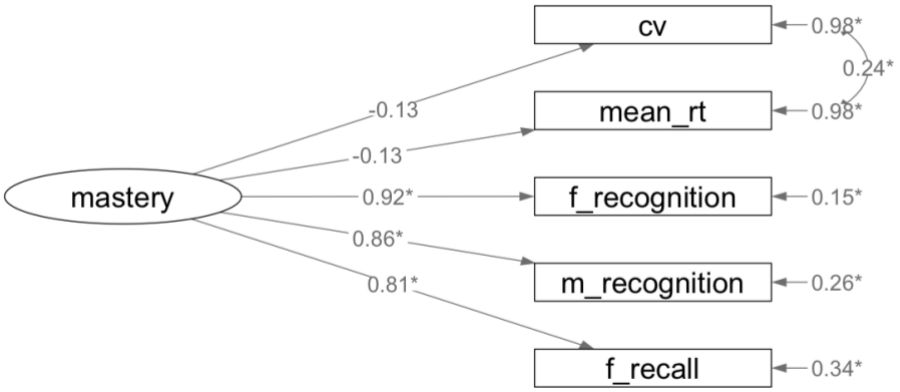

Figure 3. Confirmatory factor analysis with standardised factor loadings and error variances. Statistical significance is marked with an asterisk, with exact *p*-values and confidence intervals in Table 12 of Appendix S8.