

Supporting Information for: Dudley, A. & Marsden, E. Dimensions of lexical mastery and their relationships with listening and reading proficiency among beginner-to-low-intermediate learners of French. Article accepted in *Language Learning* on 30 April 2026.

Appendix S1: Nation's Framework of Word Knowledge and Godfroid's Expansion

Table S1.1 Godfroid's (2020, pp. 444–445) expansion of Nation's (2013, p. 49) framework with examples of different measures of lexical knowledge

Aspects of vocabulary knowledge		Receptive/productive	What does it mean to master this aspect?	Example of offline measures	Example of online measures	Fluency of use
Form	Spoken	R	What does the word sound like?	Auditory yes/no	Auditory lexical	Automaticity
			Does the learner have a formal-lexical representation (spoken form) of the new word in memory?	tests	decision tasks; auditory priming	
		P	How is the word pronounced?	Reading aloud	Naming tasks	
			How rapidly can the learner produce word forms?	exercises		
	Written	R	What does the word look like?	Proofreading	Visual lexical decision	
			Does the learner have a formal-lexical representation (written form) of the new word in memory?	exercises (identifying spelling errors)	tasks; form priming; masked repetition priming	
		P	How is the word written and spelled?	Dictation exercises	Written naming (typing)	
			How rapidly can the learner produce word forms?		tasks	
	Word parts	R	What parts are recognizable in this word?	Word segmentation	Morphological priming	
			Does the learner use morphological information to recognize a word?	tasks		
		P	What word parts are needed to express the meaning?	Generating word		
			Does the learner put together word parts to produce complex words?	derivations		

Meaning	Form and meaning	R	What meaning does this word form signal?	Meaning	Lexical decision tasks;
			How rapidly can the meaning of the word be accessed in the lexicon?	recognition/recall;	semantic categorization
		P	What word form can be used to express this meaning?	translation tasks	tasks; eye tracking
			How rapidly can the word form for a given meaning be retrieved from the lexicon?	Translation tasks	Naming tasks
	Concepts and referents	R	What is included in the concept?	Picture-based	Visual world eye-
			How rapidly can the learner access the concept and referents of the word?	vocabulary tests	tracking paradigm
		P	What items can the concept refer to?		Naming tasks
			How rapidly can the learner produce a word form for a given concept?		
	Association	R	What other words does this make us think of?	Word Associates	Semantic priming
			Has the new word been integrated into an existing semantic network in memory?	Test	
		P	What other words could we use instead of this one?	Synonyms tests	Naming tasks
			How rapidly can the learner produce word associations?		
Use	Grammatical functions	R	In what patterns does the word occur?	Grammaticality	Grammaticality
			Is the learner sensitive to ungrammatical uses of words?	judgment tasks	judgment tasks with online measures (self- paced reading, eye tracking, or event-related potentials)
		P	In what patterns must we use this word?	Guided, untimed	Free, real-time writing
			Can the learner use the word correctly in real-time conversation?	writing or speaking tasks (e.g., picture description)	or speaking tasks (e.g., story retelling)

Collocations	R	What words or types of words occur with this one? Do collocations and idioms enjoy a special status in L2 learners' lexicons?	Collocation matching tasks	Collocation priming; crosslinguistic priming;
		How are L2 idioms and L1 idioms in the lexicon related?		eye tracking
	P	What words or types of words must we use with this one?	Fill-in-the-blank exercises	Naming tasks
		Do collocations and idioms enjoy a special status in L2 learners' lexicons?		
Constraints on use	R	Where, when, and how often would we expect to meet this word?		Lexical decision tasks;
		Is the learner sensitive to how often words occur and co-occur in the language?		eye tracking
	P	Where, when, and how often can we use this word?	Guided, untimed writing or speaking tasks (e.g., picture description)	Free, real-time writing or speaking tasks (e.g., story retelling)
		Does the learner use words appropriately in context?		

Note. The four leftmost columns represent Nation's (2013) original framework, and the three rightmost columns represent Godfroid's extension. Adapted from: Table 28.2 in "Sensitive Measures of Vocabulary Knowledge and Processing" by Godfroid in *The Routledge Handbook of Vocabulary Studies* (1st ed.) edited by S. Webb. Copyright © 2020 Taylor & Francis Group. Reproduced by permission of Taylor & Francis Group.

References

- Godfroid, A. (2020). Sensitive measures of vocabulary knowledge and processing. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 433–453). Routledge.
- Nation, P. (2013). *Learning vocabulary in another language*. Cambridge University Press.

Appendix S2: Information About the General Certificate of Secondary Education in England

The General Certificate of Secondary Education (GCSE) is an academic qualification taken by most 15-to-16-year-old schoolchildren in their fifth year of secondary education in England in about eight to 11 subjects. Approximately 270,000 (of each annual cohort of approximately 600,000 students) choose to study GCSE French, German, and/or Spanish. Prior to taking these exams, students receive approximately 400-to-450 hours of classroom instruction at secondary school in the target language.

The Department for Education (2015) is responsible for stipulating the GCSE curriculum (henceforth, subject content). The subject content outlines the key knowledge and skills that the commercial awarding organizations (i.e., Assessment and Qualifications Alliance [AQA], Eduqas, and Pearson Edexcel) must test in the listening, reading, speaking, and writing exam papers. Students are entered for either Foundation or Higher tier based on their prior academic performance. Foundation tier is aimed at students expected to achieve Levels 1 to 5, and Higher tier is aimed at students expected to achieve Levels 4 to 9. Levels 4 and above are considered a pass, and Levels 5 and above a good pass.

The subject content (Department for Education, 2015) at the time of testing, which was operational for examinations taken between 2015 and 2025, explicitly outlined the grammar requirements of the qualification but not the vocabulary requirements. In other words, it did not state which lexical items test takers were expected to know in preparation for their GCSE exams. To assist with planning schemes of learning and textbooks, the awarding organizations published vocabulary lists in their exam specifications. However, prior to changes introduced during the pandemic, awarding organizations were required by the government's regulatory body, the Office of Qualifications and Examinations Regulation (Ofqual, 2021), to include words not on the awarding organizations' lists in the exams.

In 2022, the Department for Education in England announced reforms to the GCSE subject content in modern foreign languages for examinations taken from 2026 onwards. Included in these reforms was the introduction of a compulsory wordlist that awarding organizations must sample

from when creating examination papers. A requirement of the revised subject content (Department for Education, 2022, 2025) is that at least 85% of the 1,200 and 1,700 lexical items included in the wordlists at foundation and higher tier, respectively, must be high frequency (that is, sampled from the 2,000 most frequently occurring words in the target language according to a large general corpus).

References

Department for Education. (2015). *Modern foreign language: GCSE subject content*.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/485567/GCSE_subject_content_modern_foreign_langs.pdf

Department for Education. (2022/2025). *GCSE French, German and Spanish subject content*.

<https://www.gov.uk/government/publications/gcse-french-german-and-spanish-subject-content>

Ofqual. (2021). *Ofqual handbook: General conditions of recognition*.

<https://www.gov.uk/guidance/ofqual-handbook/section-d-general-requirements-for-regulated-qualifications>

Appendix S3: Critical Words

During the early stages of instructed second language development, learners' lexicons are typically very small and almost exclusively informed by the input, which, in most cases, is the language taught in the classroom. This is particularly the case for adolescent learners of foreign languages, such as French, in England, who typically receive 400-to-450 hours of exposure to the target language before sitting their high-stakes GCSE exams, with very little—if any—exposure outside the classroom. The overarching principle guiding the selection of critical words in our study was their presence in the curriculum. In this context, the curriculum was considered the GCSE wordlists published by the leading awarding organizations: AQA (2016) and Pearson Edexcel (2018). Although these wordlists were not compulsory, they were 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).

To operationalize this word selection principle, we extracted high-frequency¹ adjectives, nouns, and verbs that were common to the two leading awarding organizations' (AQA and Pearson Edexcel) vocabulary lists created for the purposes of the GCSE subject content (curriculum) at the time of testing, operational for examinations between 2015 and 2025. 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. We also ensured that the words appeared on a frequency-informed wordlist created to align with the revised GCSE subject content, operational for examinations from 2026. This decision was motivated by our long-term ambition to evaluate the impact of the revised curriculum on learners' lexical mastery.

¹ High frequency was operationalized as within the first 2,000 most frequent words in the French language, according to Lonsdale and Le Bras (2009), as per the guidance example corpus mentioned in the GCSE Subject Content (2022, 2025).

This set of words was then further checked against those that had appeared in a corpus of GCSE exam texts to ensure that the words we were testing were potentially familiar to our participants and thus stored, albeit to varying strengths and types, in their mental lexicons.

Table S3.1 Summary of critical items by frequency band and part of speech

Frequency band	Part of speech			Total
	Adjective	Noun	Verb	
1,000	5	11	10	26
2,000	5	15	4	24

Table S3.2 Breakdown of individual critical items by frequency band and part of speech

Word	Frequency Band	Part of Speech
voisin	1,000	Adjective
heureux	1,000	Adjective
étranger	1,000	Adjective
faible	1,000	Adjective
sûr	1,000	Adjective
œil	1,000	Noun
semaine	1,000	Noun
journal ^b	1,000	Noun
jeu	1,000	Noun
jour	1,000	Noun
livre	1,000	Noun
matière ^b	1,000	Noun
côté	1,000	Noun
école	1,000	Noun
femme	1,000	Noun
choix ^b	1,000	Noun
oublier	1,000	Verb
répéter ^b	1,000	Verb
aider ^b	1,000	Verb
réussir	1,000	Verb
étudier	1,000	Verb
retourner ^b	1,000	Verb
choisir ^b	1,000	Verb
naître	1,000	Verb
payer ^b	1,000	Verb
coûter ^b	1,000	Verb
utile	2,000	Adjective

égal ^b	2,000	Adjective
britannique	2,000	Adjective
sain ^b	2,000	Adjective
triste	2,000	Adjective
avril ^b	2,000	Noun
poisson ^a	2,000	Noun
professeur ^b	2,000	Noun
après-midi	2,000	Noun
mode ^b	2,000	Noun
fer	2,000	Noun
magasin ^a	2,000	Noun
examen ^b	2,000	Noun
hiver	2,000	Noun
quartier ^b	2,000	Noun
hôtel ^b	2,000	Noun
mari	2,000	Noun
fenêtre	2,000	Noun
retard ^a	2,000	Noun
lit	2,000	Noun
courir	2,000	Verb
habiter	2,000	Verb
voler	2,000	Verb
inquiéter	2,000	Verb

^a denotes a false friend, and ^b an English–French cognate.

Within the final set of 50 words, 17 (34%) were considered English–French cognates and three (6%) false friends. The inclusion of these specific words was unavoidable due to the relatively small pool of available words resulting from the overarching principles guiding word selection. At the same time, knowledge of orthographic cognates cannot be assumed, especially during the early stages of second language development, as the transparency of their meaning depends, in part, on learners' vocabulary size in their first and/or the majority language of the community, the degree of orthographic overlap, word length, and semantic similarity across the two languages.

We conducted a series of regression models to investigate any potential effect of English–French cognates and false friends on accuracy scores (in terms of form recall, meaning recognition, and form recognition) and response times. For the lexical knowledge measures, we computed

binomial logistic regression models, using the `glmer()` function from the `lme4` package in R, with item score as the outcome variable, cognate status and false friend status as ANOVA-coded fixed effects, and item and participant ID as random effects. For these analyses only, item scores for form recall were recoded as a binary variable, with responses awarded one point in the main analyses recoded as zero points, and responses awarded two points recoded as one point.

These models (Table S3.3) generally showed that whether a critical word was a cognate or a false friend did not significantly increase the probability of a participant answering an item correctly in the form-recall, meaning-recognition, and form-recognition tasks. The only exception was in the form recognition task, where a word being a false friend ($k = 3$) significantly increased the probability of a participant answering an item correctly (i.e., identifying the word as a real word). Given that only three of the 50 critical words were considered false friends, we do not consider this finding to be reliable, especially as the p -value hovered just below 0.05.

Table S3.3 Summary of the regression models investigating the effect of cognates and false friends on accuracy scores

Predictors	Form Recall			Meaning Recognition			Form Recognition		
	Odds Ratios	CI	p	Odds Ratios	CI	p	Odds Ratios	CI	p
(Intercept)	1.62	[0.61–4.31]	.335	28.16	[12.51–63.38]	< .001	50.56	[21.97–116.32]	< .001
Cognate Y–N	1.48	[0.58–3.77]	.407	1.34	[0.63–2.84]	.450	1.45	[0.68–3.12]	.337
False Friend Y–N	1.18	[0.18–7.59]	.862	3.76	[0.82–17.30]	.089	4.99	[1.02–24.33]	.047
Random Effects									
σ^2	3.29			3.29			3.29		
τ_{00}	1.89 _{id}			1.94 _{id}			1.26 _{id}		
	2.43 _{Target}			1.53 _{Target}			1.53 _{Target}		
ICC	0.57			0.51			0.46		
N	50 _{Target}			50 _{Target}			50 _{Target}		
	222 _{id}			222 _{id}			218 _{id}		
Observations	11100			11100			10678		
Marginal R^2 / Conditional R^2	.004 / .569			.015 / .521			.025 / .472		

Note. Boldface indicates statistically significant p -values (i.e., $p < .05$).

Similar analyses for response times were conducted, using the `lmer()` function from the `lme4` package in R, with response time as the outcome variable, cognate and false friend as ANOVA-coded fixed effects, and item and participant ID as random effects. Neither predictor had a significant effect on response time (Table S3.4).

Table S3.4 Summary of the regression model investigating the effect of cognates and false friends on response times

<i>Predictors</i>	Response Time		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	808.31	[758.73–857.89]	< .001
Cognate Y–N	–4.03	[–49.44–41.38]	.862
False Friend Y–N	–52.16	[–142.47–38.15]	.258
Random Effects			
σ^2	78973.37		
$\tau_{00 \text{ id}}$	15762.40		
$\tau_{00 \text{ Target}}$	5406.97		
ICC	0.21		
N_{Target}	50		
N_{id}	218		
Observations	9605		
Marginal R^2 / Conditional R^2	.002 / .213		

Note. Boldface indicates statistically significant p -values (i.e., $p < .05$).

References

AQA. (2016). *GCSE French (8658) specification*.

<https://filestore.aqa.org.uk/resources/french/specifications/AQA-8658-SP-2016.PDF>

Hawkes, R., & Lillington, C. (2016). *Viva! Edexcel GCSE (9–1) Spanish higher student book*.

Pearson Education.

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>

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>

Form recall

Word translation

You will see two lists of English words. Please type the French translation for each of the words you see. If you do not know a translation, just leave the answer blank and move on to the next word.

For instance: if you saw the English word **cat** in the list, you would type the French word **le chat** or just **chat** (either is fine) in the box next to it.

Please include accents if you know them. To insert accented letters, you can click on the corresponding button in the bar at the top.

Please do not use a dictionary. Remember: **this is not a school test, and there are no marks!** You will help us with our research only by giving honest answers. All responses are anonymous.

So, just do what you can!

Meaning recognition

Vocabulary test

Choose the French word closest to the word or phrase on the right. When you get to the bottom, click on the arrow to move on to the next screen.

Lexical decision task

You will now see some words. Some will be real French words, but others will not. For each word, decide whether it is a real French word or not.
Press the Space bar to practice.

After the practice session:

You are now ready to start the task. Try to answer as fast as possible, but still try to be accurate each time. Remember: press the left arrow for fake words, and the right arrow for real ones. Press the Space bar to begin.

Appendix S5: Form Recall Mark Scheme

Table S5.1 Mark scheme for the form recall task

Score	Criteria
2	- Target (correct translation) - Correct but different sense (e.g., <i>vivre</i> instead of <i>habiter</i> for “to live”) - Correct or target word but incorrect / missing accent - Correct or target word but incorrect / missing determiner - Correct or target but nondefault gender (e.g., <i>voisine</i> instead of <i>voisin</i>)
1	- Correct or target word but wrong form (such as plural instead of singular, or past participle instead of infinitive, e.g., <i>choisi</i> instead of <i>choisir</i>) - Correct or target word but irregular plural instead of singular (e.g., <i>yeux</i> instead of <i>œil</i>) - Correct or target root but wrong part of speech (e.g., <i>sainement</i> instead of <i>saine</i>) - Correct or target word but within a chunk, or with additional unexpected element that affects the meaning: correct word given as part of a set expression (e.g., <i>à l'étranger</i> instead of <i>étranger</i> (adjective), or <i>à la mode</i> instead of <i>mode</i>, <i>être naïtre</i> instead of <i>naître</i>)
0–1	- Correct or target translation but misspelt (score depends on how badly—give 1 if Levenshtein score $\geq .7$; give 0 if Levenshtein $< .7$)
0	- Completely nontarget (incorrect) - Blank answer - Semantically related but incorrect (e.g., <i>papier</i> for “newspaper”) - Multiword paraphrases

Special cases

- **Geographical knowledge:** Give 2 points even if they give *anglais* instead of *britannique*.
- **Misinterpretation of cue:** Still give 2 points if answer is translation of a possible interpretation of the cue (though not the intended one). For example, *bien sûr* instead of *sûr* when given “sure.”
- **Multiple cues:** If cue includes multiple words, give 2 points if answer is correct translation for at least **one** of the words in the cue (e.g., if they give *content* instead of *heureux* for “happy, lucky, fortunate,” even though *content* does not cover the meaning “lucky, fortunate”).
- **Missing auxiliary:** For example, just *né* instead of *être né* for “to be born” only gets 1 point.
- **Addition of reflexive pronouns:** Still give 2 points even if they give a reflexive pronoun that’s not needed, e.g. *s’inquiéter* instead of *inquiéter* for “to worry.”
- **Hyponyms:** Still give 2 points. For example, *roman* or *cahier* instead of *livre* for “book,” or *sprinter* or *faire du jogging* instead of *courir* for “to run.”
- **Misspellings:** Use the **Levenshtein** column in the scoring file to determine whether to give misspellings a 0 or 1 score. Sometimes, participants gave a correct but nontarget answer and

didn't get the spelling right (e.g., the target answer was *livre*, so *cahier* would be a correct answer, but the participant wrote *cachier*). In that case, disregard the Levenshtein column in the scoring file (which does not apply in this case because it's for the similarity between *livre* and *cachier*) and use the Levenshtein score calculator to determine the similarity between *cachier* and *cahier*.

- **Multiple answers:** Score based on the best answer.

Appendix S6: Effect of Year of Data Collection on Lexical Knowledge and Processing

To reach the minimum sample size required for confirmatory factor analyses, data collection was conducted across two years (2022 and 2023). We, therefore, ran a series of linear regression models using the built-in `lm()` function in R, with overall accuracy, the coefficient of variation, or mean response time as the outcome variable and year as an ANOVA-coded fixed effect in order to examine the extent to which year of data collection influenced performance on the measures of lexical knowledge, lexical processing, and reading and listening comprehension. These models (Tables S6.1 to S6.3) consistently showed that year of data collection had no effect on overall accuracy, regardless of measure, the coefficient of variation, or mean response time.

Table S6.1 Summary of the regression models investigating the effect of year of data collection on lexical knowledge

Predictors	Form Recall			Form Recognition			Meaning Recognition		
	Estimates	CI	<i>p</i>	Estimates	CI	<i>p</i>	Estimates	CI	<i>p</i>
(Intercept)	0.00	−0.13–0.14	.962	0.00	−0.13–0.13	.997	0.00	−0.13–0.14	.981
2023 vs. 2022	0.23	−0.03–0.50	.083	0.02	−0.25–0.29	.892	0.12	−0.15–0.38	.393
Observations	218			218			218		
R^2 / R^2 adjusted	0.014 / 0.009			0.000 / −0.005			0.003 / −0.001		

Table S6.2 Summary of the regression models investigating the effect of year of data collection on lexical processing

Predictors	Coefficient of Variation			Mean Response Time		
	Estimates	CI	<i>p</i>	Estimates	CI	<i>p</i>
(Intercept)	−0.00	−0.14–0.13	.979	−0.00	−0.13–0.13	.995
2023 vs. 2022	−0.13	−0.40–0.14	.344	−0.03	−0.30–0.24	.813
Observations	218			218		
R^2 / R^2 adjusted	0.004 / −0.000			0.000 / −0.004		

Table S6.3 Summary of the regression models investigating the effect of year of data collection on Diplôme d'Études en Langue Française (“Diploma in French Language Studies”; DELF) listening and reading

Predictors	DELF listening			DELF reading		
	Estimates	CI	<i>p</i>	Estimates	CI	<i>p</i>
(Intercept)	−0.00	−0.14–0.14	1.000	0.00	−0.13–0.13	.997
2023 vs. 2022	−0.08	−0.35–0.19	.546	0.02	−0.25–0.29	.862
Observations	212			216		
R^2 / R^2 adjusted	0.002 / −0.003			0.000 / −0.005		

Appendix S7: Spearman's Correlations With 95% Confidence Intervals

Table S7.1 Spearman's correlations (with 95% confidence intervals) between knowledge, processing, and proficiency measures

Comparison	<i>rho</i> [95% CI]	<i>p</i> -value
Form recall–Meaning recognition	.774 [.714, .822]	< .001
Form recall–Form recognition (raw)	.697 [.622, .760]	< .001
Form recall–Form recognition (adjusted)	.745 [.679, .799]	< .001
Form recall–Mean RT	–.123 [–.252, .010]	.070
Form recall–CV	–.085 [–.215, .048]	.211
Form recall–DELFL listening	.580 [.483, .663]	< .001
Form recall–DELFL reading	.604 [.512, .683]	< .001
Form recall–GCSE listening	.626 [.532, .705]	< .001
Form recall–GCSE reading	.595 [.495, .679]	< .001
Meaning recognition–Form recognition (raw)	.753 [.689, .805]	< .001
Meaning recognition–Form recognition (adjusted)	.822 [.773, .861]	< .001
Meaning recognition–Mean RT	–.131 [–.259, .002]	.053
Meaning recognition–CV	–.105 [–.235, .028]	.122
Meaning recognition–DELFL listening	.688 [.610, .753]	< .001
Meaning recognition–DELFL reading	.711 [.638, .771]	< .001
Meaning recognition–GCSE listening	.705 [.626, .770]	< .001
Meaning recognition–GCSE reading	.709 [.631, .773]	< .001
Form recognition (raw)–Form recognition (adjusted)	.911 [.885, .931]	< .001
Form recognition (raw)–Mean RT	–.137 [–.265, –.004]	.044
Form recognition (raw)–CV	–.032 [–.164, .101]	.640
Form recognition (raw)–DELFL listening	.645 [.559, .718]	< .001
Form recognition (raw)–DELFL reading	.652 [.568, .723]	< .001
Form recognition (raw)–GCSE listening	.698 [.618, .764]	< .001
Form recognition (raw)–GCSE reading	.679 [.595, .749]	< .001
Form recognition (adjusted)–Mean RT	–.100 [–.230, .033]	.141
Form recognition (adjusted)–CV	–.077 [–.208, .056]	.255
Form recognition (adjusted)–DELFL listening	.666 [.584, .735]	< .001
Form recognition (adjusted)–DELFL reading	.704 [.630, .766]	< .001
Form recognition (adjusted)–GCSE listening	.718 [.642, .780]	< .001
Form recognition (adjusted)–GCSE reading	.752 [.683, .807]	< .001
Mean RT–CV	.365 [.244, .475]	< .001
Mean RT–DELFL listening	–.065 [–.198, .071]	.350
Mean RT–DELFL reading	–.066 [–.198, .068]	.331
Mean RT–GCSE listening	–.118 [–.255, .024]	.103
Mean RT–GCSE reading	–.104 [–.242, .037]	.149
CV–DELFL listening	–.087 [–.220, .048]	.205
CV–DELFL reading	–.118 [–.247, .016]	.085
CV–GCSE listening	–.077 [–.215, .065]	.290
CV–GCSE reading	–.099 [–.237, .043]	.172
DELFL listening–DELFL reading	.717 [.644, .777]	< .001
DELFL listening–GCSE listening	.733 [.660, .793]	< .001
DELFL listening–GCSE reading	.633 [.539, .712]	< .001
DELFL reading–GCSE listening	.697 [.615, .763]	< .001
DELFL reading–GCSE reading	.664 [.576, .736]	< .001
GCSE listening–GCSE reading	.791 [.732, .839]	< .001

Note. GCSE = General Certificate of Secondary Education; DELF = Diplôme d'Études en Langue Française (“Diploma in French Language Studies”; DELF); RT = response time; CV = coefficient of variation. Boldface indicates statistically significant p -values (i.e., $p < .05$).

Appendix S8: Confirmatory Factor Analyses

Table S8.1 Standardized estimates, squared multiple correlations, and *z*-values for the one-factor model

Path	Standardized Estimates [95% CIs]	<i>SE</i>	<i>R</i> ²	<i>z</i>	<i>p</i>
Lexical Mastery =~ Form recall	.836 [.776, .896]	.031	.699	27.188	< .001
Lexical Mastery =~ Meaning recognition	.860 [.809, .911]	.026	.740	32.998	< .001
Lexical Mastery =~ Form recognition	.921 [.880, .963]	.021	.848	43.530	< .001
Lexical Mastery =~ Mean response time	−.123 [−.245, −.001]	.062	.015	−1.979	.048
Lexical Mastery =~ Coefficient of variation	−.134 [−.291, .023]	.080	.018	−1.669	.095
Mean response time ~~ Coefficient of variation	.238 [.052, .425]	.095	—	2.501	.012

Note. =~ indicates a loading; ~~ indicates a correlation. Boldface indicates statistically significant *p*-values (i.e., *p* < .05).

Appendix S9: Hierarchical Linear (for Diplôme d'Études en Langue Française; DELF) and Ordinal
(for General Certificate of Secondary Education; GCSE) Regression Models

Table S9.1 Knowledge and processing as predictors of comprehension

Steps	Predictors	(pseudo) R^2	95% CI	R^2_{adjusted}	ΔR^2	p
<i>DELF listening comprehension</i>						
1	Form recognition, meaning recognition, form recall	.460	[.364, .556]	.452		
2	Form recognition, meaning recognition, form recall, CV	.460	[.364, .556]	.450	.000	.876
2	Form recognition, meaning recognition, form recall, RT	.460	[.364, .556]	.450	.000	.875
<i>GCSE listening comprehension</i>						
1	Form recognition, meaning recognition, form recall	.533				
2	Form recognition, meaning recognition, form recall, CV	.535			.002	.493
2	Form recognition, meaning recognition, form recall, RT	.535			.002	.486
<i>DELF reading comprehension</i>						
1	Form recognition, meaning recognition, form recall	.549	[.462, .636]	.543		
2	Form recognition, meaning recognition, form recall, CV	.550	[.464, .637]	.542	.001	.454
2	Form recognition, meaning recognition, form recall, RT	.549	[.463, .636]	.541	.000	.681
<i>GCSE reading comprehension</i>						
1	Form recognition, meaning recognition, form recall	.567				
2	Form recognition, meaning recognition, form recall, CV	.567			.000	.882
2	Form recognition, meaning recognition, form recall, RT	.569			.002	.472

Note. To the best of our knowledge, it is not possible to compute 95% confidence intervals (CIs) around pseudo R^2 (necessitated by the ordinal nature of the GCSE data). DELF = Diplôme d'Études en Langue Française “Diploma in French Language Studies”; GCSE = General Certificate of Secondary Education; CV = coefficient of variation; RT = response time.

Table S9.2 Summary of the Diplôme d'Études en Langue Française (DELF; “Diploma in French Language Studies”) listening regression model (knowledge only)

DELF listening						
Predictors	Estimate	95% CI	p	RelImp	95% CI	Rank
(Intercept)	−0.00	[−0.10–0.10]	.994			
Form recognition	0.39	[0.20–0.57]	< . .001	18.86	[13.25, 25.32]	1
Meaning recognition	0.22	[0.05–0.39]	.011	14.93	[11.31, 19.02]	2
Form recall	0.12	[−0.05–0.29]	.156	12.21	[8.26, 17.05]	3
Observations	212					
R^2 / R^2 adjusted	.460 / .452					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant p -values (i.e., $p < .05$).

Table S9.3 Summary of the Diplôme d’Études en Langue Française (DELF; “Diploma in French Language Studies”) listening regression model (knowledge and coefficient of variation)

DELF listening						
Predictors	Estimate	95% CI	<i>p</i>	RelImp	95% CI	Rank
(Intercept)	−0.00	[−0.10–0.10]	.990			
Form recognition	0.39	[0.20–0.57]	< .001	18.73	[13.19, 24.94]	1
Meaning recognition	0.22	[0.05–0.39]	.013	14.77	[11.00, 18.75]	2
Form recall	0.12	[−0.05–0.29]	.155	12.14	[8.19, 16.73]	3
Coefficient of variation	−0.01	[−0.11–0.10]	.876	0.36	[0.06, 2.40]	4
Observations	212					
<i>R</i> ² / <i>R</i> ² adjusted	.460 / .450					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., *p* < .05).

Table S9.4 Summary of the Diplôme d’Études en Langue Française (DELF; “Diploma in French Language Studies”) listening regression model (knowledge and mean response time)

DELF listening						
Predictors	Estimate	95% CI	<i>p</i>	RelImp	95% CI	Rank
(Intercept)	−0.00	[−0.10–0.10]	.994			
Form recognition	0.39	[0.20–0.57]	< .001	18.78	[13.29, 24.80]	1
Meaning recognition	0.22	[0.05–0.39]	.011	14.86	[11.29, 19.46]	2
Form recall	0.12	[−0.05–0.29]	.159	12.14	[8.10, 17.11]	3
Mean response time	−0.01	[−0.11–0.09]	.875	0.22	[0.03, 1.70]	4
Observations	212					
<i>R</i> ² / <i>R</i> ² adjusted	.460 / .450					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., *p* < .05).

Table S9.5 Summary of the Diplôme d’Études en Langue Française (DELF; “Diploma in French Language Studies”) reading regression model (knowledge only)

DELF reading						
Predictors	Estimate	95% CI	<i>p</i>	RelImp	95% CI	Rank
(Intercept)	−0.00	[−0.09–0.09]	.985			
Form recognition	0.41	[0.24–0.58]	< .001	22.12	[16.47, 28.26]	1
Meaning recognition	0.32	[0.16–0.47]	< .001	19.87	[15.61, 24.54]	2
Form recall	0.06	[−0.09–0.21]	.447	12.92	[8.58, 18.16]	3
Observations	216					
<i>R</i> ² / <i>R</i> ² adjusted	.549 / .543					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., *p* < .05).

Table S9.6 Summary of the Diplôme d'Études en Langue Française (DELF; “Diploma in French Language Studies”) reading regression model (knowledge and coefficient of variation)

DELF reading						
Predictors	Estimate	95% CI	<i>p</i>	RelImp	95% CI	Rank
(Intercept)	−0.00	[−0.09–0.09]	.985			
Form recognition	0.41	[0.24–0.58]	< . .001	21.92	[16.28, 28.18]	1
Meaning recognition	0.31	[0.15–0.47]	< . .001	19.56	[15.36, 24.65]	2
Form recall	0.06	[−0.09–0.21]	.414	12.85	[8.63, 18.12]	3
Coefficient of variation	−0.04	[−0.13–0.06]	.454	0.69	[0.07, 3.39]	4
Observations	216					
<i>R</i> ² / <i>R</i> ² adjusted	.550 / .542					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., *p* < .05).

Table S9.7 Summary of the Diplôme d'Études en Langue Française (DELF; “Diploma in French Language Studies”) reading regression model (knowledge and mean response time)

DELF reading						
Predictors	Estimate	95% CI	<i>p</i>	RelImp	95% CI	Rank
(Intercept)	−0.00	[−0.09–0.09]	.988			
Form recognition	0.41	[0.24–0.58]	< . .001	22.10	[16.72, 28.34]	1
Meaning recognition	0.32	[0.16–0.47]	< . .001	19.84	[15.49, 24.61]	2
Form recall	0.06	[−0.09–0.21]	.437	12.89	[8.65, 18.28]	3
Mean response time	0.02	[−0.07–0.11]	.681	0.12	[0.03, 1.25]	4
Observations	216					
<i>R</i> ² / <i>R</i> ² adjusted	.549 / .541					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., *p* < .05).

Table S9.8 Summary of the General Certificate of Secondary Education (GCSE) listening regression model (knowledge only)

GCSE listening					
Predictors	Odds Ratios	95% CI	<i>p</i>	RelImp	Rank
Form recognition	3.45	[2.04–5.93]	< . .001	22.70	1
Meaning recognition	1.41	[0.87–2.38]	.173	15.30	3
Form recall	1.62	[1.04–2.54]	.036	15.35	2
Observations	193				
Nagelkerke's <i>R</i> ²	.533				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo *R*²). Boldface indicates statistically significant *p*-values (i.e., *p* < .05).

Table S9.9 Summary of the General Certificate of Secondary Education (GCSE) listening regression model (knowledge with coefficient of variation)

GCSE listening					
Predictors	Odds Ratios	CI	<i>p</i>	RelImp	Rank
Form recognition	3.48	[2.05–5.98]	< .001	22.66	1
Meaning recognition	1.44	[0.89–2.41]	.151	15.27	= 2
Form recall	1.60	[1.03–2.51]	.040	15.27	= 2
Coefficient of variation	1.10	[0.84–1.45]	.494	0.26	4
Observations	193				
Nagelkerke's R^2	.535				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Table S9.10 Summary of the General Certificate of Secondary Education (GCSE) listening regression model (knowledge with RT)

GCSE listening					
Predictors	Odds Ratios	CI	<i>p</i>	RelImp	Rank
Form recognition	3.40	[2.00–5.86]	< .001	22.46	1
Meaning recognition	1.42	[0.88–2.39]	.170	15.20	3
Form recall	1.62	[1.04–2.54]	.036	15.23	2
Mean response time	0.90	[0.66–1.22]	.487	0.58	4
Observations	193				
Nagelkerke's R^2	.535				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Table S9.11 Summary of the General Certificate of Secondary Education (GCSE) reading regression model (knowledge only)

GCSE reading					
Predictors	Odds Ratios	95% CI	<i>p</i>	RelImp	Rank
Form recognition	4.12	[2.39–7.30]	< .001	24.92	1
Meaning recognition	2.19	[1.26–4.04]	.009	19.08	2
Form recall	1.04	[0.64–1.70]	.860	12.74	3
Observations	193				
Nagelkerke's R^2	.567				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Table S9.12 Summary of the General Certificate of Secondary Education (GCSE) reading regression model (knowledge with coefficient of variation)

GCSE reading					
Predictors	Odds Ratios	CI	<i>p</i>	RelImp	Rank
Form recognition	4.12	[2.39 – 7.31]	< .001	24.78	1
Meaning recognition	2.18	[1.25 – 4.04]	.010	18.92	2
Form recall	1.05	[0.64 – 1.70]	.856	12.67	3
Coefficient of variation	0.98	[0.73 – 1.31]	.882	0.38	4
Observations	193				
Nagelkerke's R^2	.567				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Table S9.13 Summary of the General Certificate of Secondary Education (GCSE) reading regression model (knowledge with RT)

GCSE reading					
Predictors	Odds Ratios	CI	<i>p</i>	RelImp	Rank
Form recognition	4.11	[2.37–7.30]	< .001	24.77	1
Meaning recognition	2.20	[1.26–4.07]	.009	19.00	2
Form recall	1.04	[0.64–1.69]	.865	12.67	3
Mean response time	0.89	[0.64–1.23]	.473	0.43	4
Observations	193				
Nagelkerke's R^2	.569				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Appendix S10: Meaning Recall Analyses

In addition to the measures of lexical knowledge reported in the main body of the manuscript, a subset of participants ($n = 87$) completed a test of meaning recall approximately two weeks after taking part in the main experiment (to reduce any possible priming effects). In this test, participants were presented with a list of 60 high-frequency words in French and were asked to translate these words into English—the majority language of the participants. These 60 words were randomized across participants and included 30 of the critical words sampled from the main study (and thus the participants’ curriculum). The remaining 30 were high-frequency words not featured in the main study or the participants’ school curriculum. Table S10.1 below presents an overview of the critical words.

Table S10.1 List of critical words according to frequency band and part of speech

Word	Frequency Band	Part of Speech
étranger	1000	Adjective
faible	1000	Adjective
heureux	1000	Adjective
voisin	1000	Adjective
école	1000	Noun
femme	1000	Noun
jeu	1000	Noun
jour	1000	Noun
journal	1000	Noun
matière	1000	Noun
semaine	1000	Noun
aider	1000	Verb
étudier	1000	Verb
oublier	1000	Verb
réussir	1000	Verb
britannique	2000	Adjective
triste	2000	Adjective
utile	2000	Adjective
après-midi	2000	Noun

examen	2000	Noun
fenêtre	2000	Noun
hiver	2000	Noun
magasin	2000	Noun
poisson	2000	Noun
professeur	2000	Noun
retard	2000	Noun
courir	2000	Verb
habiter	2000	Verb
inquiéter	2000	Verb
voler	2000	Verb

In this appendix, we report on learners' knowledge of the aforementioned 30 critical words across four types: form recognition, meaning recognition, form recall, and meaning recall. For a description of the first three measures, see Vocabulary Measures in the main manuscript. In particular, we seek to understand (a) the degree of interrelatedness between these four types of lexical knowledge, using (Spearman's) correlations², and (b) the relative importance of these four types in predicting listening and reading comprehension, using dominance analyses from regression modelling.

² Unlike the main study presented in the manuscript, we did not conduct confirmatory factor analyses in order to examine the unidimensionality of lexical knowledge as a construct given the relatively small sample size. Neither do we report on the relations between lexical knowledge and lexical processing, given the lack of correlations between the three types of lexical knowledge (form recognition, meaning recognition, and form recall) and the two types of lexical processing (mean RT and CV) reported as part of the main study.

Descriptive Statistics

Table S10.2 Descriptive statistics and internal consistency reliability for the knowledge measures

	Form Recall	Meaning Recall	Form Recognition (Raw)	Form Recognition (Adjusted ¹)	Meaning Recognition
<i>n</i>	87	87	84	84	87
Mean	66.42%	80.92%	92.09%	72.03%	87.89%
<i>SD</i>	18.92%	17.61%	8.27%	19.19%	13.11%
Median	66.67%	86.67%	93.33%	74.55%	90.00%
95% CI	[62.39%, 70.45%]	[77.17%, 84.67%]	[90.30%, 93.89%]	[67.87%, 76.20%]	[85.10%, 90.69%]
Min	21.67%	26.67%	65.52%	22.93%	26.67%
Max	100.00%	100.00%	100.00%	100.00%	100.00%
Skew	−0.43	−1.11	−1.31	−0.57	−2.06
Kurtosis	−0.53	0.64	1.19	−0.45	5.99
Omega	.90	.91	.75	—	.90
Alpha	.89	.89	.72	—	.87

¹ I_{SDT} = index of signal detection.

The internal consistency reliability was generally good-to-excellent for the knowledge measures (Table S10.2). Descriptive statistics demonstrate that learners knew, on average, at least half of the words in each of the knowledge measures and that meaning recognition was consistently the strongest, followed by meaning recall, form recognition (once corrected for false alarm rates), and then form recall, with mostly nonoverlapping CIs between pairs of measures. The range in accuracy scores across the knowledge measures indicates that our approach to word selection appropriately captured a range of words that learners knew to varying degrees. Henceforth, only the adjusted (not the raw) form recognition scores will be used, as these scores factor in the effect of false alarm rates (i.e., guessing behavior).

Table S10.3 Descriptive statistics and internal consistency reliability for the Diplôme d'Études en Langue Française (DELF; Diploma in French Language Studies) proficiency measures for a subset of learners who also undertook a meaning-recall test

	Listening	Reading
<i>n</i>	87	87
Mean (%)	51.68%	71.26%
SD (%)	22.40%	24.53%
Median (%)	52.00%	76.00%
95% CI (%)	[46.90%, 56.45%]	[66.04%, 76.49%]
Min (%)	16.00%	12.00%
Max (%)	96.00%	100.00%
Skew	0.19	-0.60
Kurtosis	-1.06	-0.83
Omega	.85	.90
Alpha	.85	.90

Table S10.4 Descriptive statistics for French General Certificate of Secondary Education (GCSE) level

	Percentage of learners achieving each level								Total ¹
	U	3	4	5	6	7	8	9	
Listening	2.70%	4.05%	5.41%	21.62%	4.05%	16.22%	25.68%	20.27%	74
Reading	4.05%	5.41%	8.11%	13.51%	6.76%	6.76%	14.86%	40.54%	74

¹ Not all participants sent an individual breakdown of their GCSE results.

Table S10.5 Spearman’s correlational analyses between the knowledge and proficiency measures

	Form Recall	Meaning Recall	Meaning Recognition	Form Recognition (Raw)	Form Recognition (Adjusted)	DELFL Listening	DELFL Reading	GCSE Listening
Meaning Recall	.768 (< .001)							
Meaning Recognition	.765 (< .001)	.807 (< .001)						
Form Recognition (Raw)	.641 (< .001)	.728 (< .001)	.701 (< .001)					
Form Recognition (Adjusted)	.657 (< .001)	.741 (< .001)	.737 (< .001)	.921 (< .001)				
DELFL Listening	.582 (< .001)	.712 (< .001)	.746 (< .001)	.669 (< .001)	.695 (< .001)			
DELFL Reading	.593 (< .001)	.764 (< .001)	.712 (< .001)	.701 (< .001)	.694 (< .001)	.798 (< .001)		
GCSE Listening	.607 (< .001)	.777 (< .001)	.712 (< .001)	.613 (< .001)	.648 (< .001)	.789 (< .001)	.710 (< .001)	
GCSE Reading	.531 (< .001)	.710 (< .001)	.682 (< .001)	.604 (< .001)	.656 (< .001)	.703 (< .001)	.629 (< .001)	.846 (< .001)

Computed correlation used Spearman method with pairwise deletion.

Note. *p*-values are presented in parentheses in the table above. Boldface indicates statistically significant *p*-values (i.e., $p < .05$). DELF = Diplôme d’Études en Langue Française (“Diploma in French Language Studies”); GCSE = General Certificate of Secondary Education.

Relationship between knowledge measures. Spearman’s correlations (Table S10.5) revealed a high degree of interrelatedness ($r \geq .60$; Plonsky & Oswald, 2014) between the four knowledge measures (form recall, meaning recall, meaning recognition, form recognition), ranging from .641 to .807, but not so strong to raise concerns about multicollinearity ($r \geq .90$; Kline, 2023, p. 56).

Relationship between knowledge and proficiency measures. All four knowledge measures showed highly significant medium-to-large positive correlations with listening measures, ranging from .582 to .777, and with reading measures, ranging from .531 to .764 (Table S10.5).

Summary. In sum, we found a very high degree of interrelatedness both within the knowledge measures and between the knowledge and proficiency measures, regardless of modality. The findings from this substudy are thus in line with those reported in the main study.

Relative importance of vocabulary knowledge to listening and reading comprehension

A series of regression models were computed to explore the extent to which the four types of lexical knowledge contributed to L2 listening and reading comprehension. As per the main study, we used dominance weights to ascertain each predictor's relative importance. For further discussion of this decision, see Analyses in the main body of the article. Note that no evidence of multicollinearity was found between predictor variables ($VIF < 5$ for each model). We applied Box-Cox transformations to the DELF reading data to meet the normality and homoscedasticity assumptions for the linear models and merged Levels 4³ and below into one level to meet the proportional odds assumption for the GCSE ordinal models.

Listening models. Lexical knowledge as a global construct—combining form recall, meaning recall, form recognition, and meaning recognition—explained 54.11% (95% CI [40.77%, 67.46%]) of the variance in the DELF listening model (Table S10.6). Of these four predictors, form recognition explained the most variance (as per the main study), followed by meaning recognition, meaning recall, and then form recall.

Table S10.6 Summary of the Diplôme d'Études en Langue Française (DELF; "Diploma in French Language Studies") listening regression model (knowledge only)

DELF listening						
Predictors	Estimate	95% CI	<i>p</i>	<i>RelImp</i>	95% CI	Rank
(Intercept)	−0.00	[−0.15–0.15]	.994			
Form recognition	0.39	[0.15–0.64]	.002	18.38	[10.85, 26.79]	1
Meaning recognition	0.28	[−0.04–0.61]	.088	14.04	[9.49, 23.00]	2
Form recall	−0.09	[−0.37–0.20]	.548	8.45	[5.26, 12.98]	4
Meaning recall	0.26	[−0.06–0.57]	.111	13.25	[8.63, 20.10]	3
Observations	84					
R^2 / R^2 adjusted	.541 / .518					

Note. CI = Confidence interval, *RelImp* = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

³ Level 4 or above is considered a pass at GCSE.

A slightly different pattern of results emerged for the GCSE listening model (Table S10.7). Although the four knowledge types explained 58.32% of the variance in the GCSE listening model, a different order of predictor importance emerged. Specifically, we found that meaning recall explained the most variance, followed by form recognition, meaning recognition, and then form recall.

Table S10.7 Summary of the General Certificate of Secondary Education (GCSE) listening regression model (knowledge only)

GCSE listening					
Predictors	Odds Ratios	95% CI	<i>p</i>	RelImp	Rank
Form recognition	2.18	[1.07–4.54]	.038	13.88	2
Meaning recognition	1.18	[0.49–3.18]	.723	12.40	3
Form recall	0.90	[0.42–1.93]	.793	10.07	4
Meaning recall	5.12	[1.99–14.30]	.002	21.96	1
Observations	73				
Nagelkerke's R^2	.583				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Reading models. Lexical knowledge predicted 60.38% (95% CI [48.20%, 72.55%]) of the variance in the DELF reading model (Table S10.8). In line with the GCSE listening model, we found that meaning recall explained the most variance, followed by form recognition, meaning recognition, and form recall.

Table S10.8 Summary of the Diplôme d'Études en Langue Française (DELF; "Diploma in French Language Studies") reading regression model (knowledge only)

DELF reading						
<i>Predictors</i>	<i>Estimate</i>	<i>95% CI</i>	<i>p</i>	<i>RelImp</i>	<i>95% CI</i>	<i>Rank</i>
(Intercept)	−0.01	[−0.15–0.13]	.894			
Form recognition	0.33	[0.10–0.55]	.005	17.20	[10.55, 24.98]	2
Meaning recognition	0.19	[−0.12–0.49]	.222	13.74	[9.17, 21.33]	3
Form recall	−0.12	[−0.38–0.14]	.352	9.67	[5.82, 16.31]	4
Meaning recall	0.49	[0.20–0.79]	.001	19.77	[13.51, 28.00]	1
Observations	84					
R^2 / R^2 adjusted	.604 / .584					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant p -values (i.e., $p < .05$).

Although lexical knowledge (as a global construct) predicted 55.58% of the variance in the GCSE reading model (Table S10.9), a slightly different order of importance emerged, relative to the DELF reading model: Meaning recall explained the most variance, followed very closely by meaning recognition, form recognition, and then form recall. These findings are thus largely consistent with those reported in the main study.

Table S10.9 Summary of the General Certificate of Secondary Education (GCSE) reading regression model (knowledge only)

Predictors	GCSE reading				
	Odds Ratios	95% CI	p	RelImp	Rank
Form recognition	1.50	[0.72–3.19]	.283	13.98	3
Meaning recognition	3.70	[1.10–13.83]	.050	16.42	2
Form recall	0.77	[0.32–1.81]	.547	8.50	4
Meaning recall	3.20	[1.22–9.04]	.025	16.69	1
Observations	73				
Nagelkerke's R^2	.556				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant p -values (i.e., $p < .05$).

Summary

In sum, the findings reported in this substudy closely mirror those reported in the main study concerning the relative importance of form recall, form recognition, and meaning recognition in explaining reading and listening proficiency. In three of the four models (GCSE listening and DELF and GCSE reading) reported in this substudy, we found that meaning recall was the strongest (or joint strongest) predictor of comprehension (but the second weakest predictor in the DELF listening model). The importance of meaning recall, especially in reading comprehension, thus aligns with the order of importance observed in Zhang and Zhang's (2022) meta-analysis on the contribution of different types of lexical knowledge to listening and reading comprehension.

References

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

Appendix S11: Alternative Analyses for Research Questions 1 and 2 Using Huibregtse et al.'s

(2002) Formula

As noted in the main text, there are two formulae in circulation for the index of signal detection (which corrects form-recognition accuracy rates for guessing behavior). In the Results section of the main text, we presented analyses which used the formula reported in Zhang et al. (2020). In this appendix, we present alternative analyses which use the following formula, as reported in Huibregtse et al. (2002):

$$1 - \frac{4h(1 - f) - 2(h - f)(1 + h - f)}{4h(1 - f) - (h - f)(1 + h - f)}$$

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

In general, a similar pattern of results pertained, regardless of which formula was used. There were, however, two relatively small differences in findings, which we summarize here. First, mean form-recognition accuracy (once adjusted for guessing behavior) was higher by 10.03 percentage points (75.00% versus 64.97%) when we used the formula reported in Huibregtse et al. (2002) rather than the formula reported in Zhang et al. (2020). Second, the amount of variance explained by the three knowledge measures was, on average, 2.61 percentage points ($SD = 0.88$ percentage points) lower in the four models of listening and reading proficiency when we used the formula reported in Huibregtse et al. (2002) rather than the formula reported in Zhang et al. (2020).

Descriptive Statistics

Internal consistency reliability was generally good-to-excellent for the three knowledge measures (Table S11.1 and Table S11.2). Descriptive statistics demonstrate that learners knew, on average, 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 nonoverlapping 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 behavior.

Table S11.1 Descriptive statistics and internal consistency reliability for the knowledge measures

	Form recall	Meaning recognition	Form recognition (Raw)	Form recognition (Adjusted ¹)
<i>n</i>	218	218	218	218
Mean (%)	61.70%	85.63%	89.86%	75.00%
<i>SD</i> (%)	20.18%	13.75%	9.14%	15.52%
Median (%)	63.00%	90.00%	92.00%	78.37%
95% CI (%)	[59.00%, 64.39%]	[83.80%, 87.47%]	[88.64%, 91.08%]	[72.93%, 77.08%]
Min (%)	9.00%	28.00%	54.55%	34.59%
Max (%)	99.00%	100.00%	100.00%	100.00%
Skew	−0.41	−1.63	−1.37	−0.82
Kurtosis	−0.37	3.10	1.87	−0.16
Omega	.95	.93	.83	—
Alpha	.94	.92	.82	—

¹ I_{SDT} = index of signal detection.

Table S11.2 Descriptive statistics and internal consistency reliability for the processing measures

	Mean Response Time (in ms)	Coefficient of Variation
<i>n</i>	218	218
Mean	824.32	0.33
<i>SD</i>	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 response time and coefficient of variation were based only on response times for hits (correct yes responses on the form recognition test).

Table S11.3 Descriptive statistics and internal consistency reliability for the Diplôme d'Études en Langue Française (DELF; Diploma in French Language Studies) proficiency measures

	Listening	Reading
<i>n</i>	212	216
Mean	53.57%	71.69%
Median	52.00%	76.00%
<i>SD</i>	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 did not complete more than half of each test or achieved scores below 10% ($n = 6$ for the listening test; $n = 2$ for the reading test) were excluded from the dataset, as such low performance or test completion may indicate lack of engagement in the test.

Table S11.4 Descriptive statistics for French General Certificate of Secondary Education (GCSE) level

	Percentage of learners achieving each level								Total ¹
	U	3	4	5	6	7	8	9	
Reading	2.07%	4.15%	4.15%	14.00%	5.18%	7.77%	18.60%	44.00%	193
Listening	2.59%	2.07%	5.70%	17.10%	4.15%	20.70%	19.70%	28.00%	193

¹ 88.53% (193) of the 218 participants sent an individual breakdown of their GCSE results.

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 ρ (.365) indicated a small (bordering on medium) statistically significant, positive relationship between mean RT and CV (Table S11.5). Thus, we refer to CV as a marker of processing stability and automaticity, henceforth.

Table S11.5 Spearman’s correlations between the knowledge, processing, and proficiency measures.

	Form recall	Meaning recognition	Form recognition (raw)	Form recognition (adjusted)	Mean response time	Coefficient of variation	DELFL list.	DELFL read.	GCSE list.
Meaning recognition	.774 (< .001)								
Form recognition (raw)	.697 (< .001)	.753 (< .001)							
Form recognition (adjusted)	.696 (< .001)	.779 (< .001)	.745 (< .001)						
Mean response time	-.123 (.070)	-.131 (.053)	-.137 (.044)	-.069 (.311)					
Coefficient of variation	-.085 (.211)	-.105 (.122)	-.032 (.640)	-.124 (.067)	.365 (< .001)				
DELFL listening	.580 (< .001)	.688 (< .001)	.645 (< .001)	.614 (< .001)	-.065 (.350)	-.087 (.205)			
DELFL reading	.604 (< .001)	.711 (< .001)	.652 (< .001)	.671 (< .001)	-.066 (.331)	-.118 (.085)	.717 (< .001)		
GCSE listening	.626 (< .001)	.705 (< .001)	.698 (< .001)	.659 (< .001)	-.118 (.103)	-.077 (.290)	.733 (< .001)	.697 (< .001)	
GCSE reading	.595 (< .001)	.709 (< .001)	.679 (< .001)	.727 (< .001)	-.104 (.149)	-.099 (.172)	.633 (< .001)	.664 (< .001)	.791 (< .001)

Note. *p*-values are presented in parentheses in the table above. Boldface indicates statistically significant *p*-values (i.e., $p < .05$). Small: $r > .25$; medium: $r > .40$; large: $r > .60$ (Plonsky & Oswald, 2014). DELFL = Diplôme d’Études en Langue Française (“Diploma in French Language Studies”); GCSE = General Certificate of Secondary Education.

Relationships Between Knowledge and Processing Measures

Spearman’s correlations (Table S11.5) revealed a high degree of interrelatedness ($r \geq .60$, Plonsky & Oswald, 2014) between the knowledge measures (form recall, meaning recognition, form recognition), ranging from .696 to .779, but not so strong to raise concerns about multicollinearity ($r \geq .90$; Kline, 2023, p. 56). 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 ($\rho = -.137$). 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, as suggested by an anonymous reviewer. 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 .580 to .705, and reading measures, ranging from .595 to .727 (Table S11.5). However, there was no significant 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.

Research Question 1: 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. The nonindependence of mean RT and CV was accounted for by allowing the two processing measures to covary.

The overall fit indices (Table S11.6) suggested that the one-factor model demonstrated a good overall fit: SRMR and CFI—our focus here, given the relatively small sample size (Kenny, 2024)—were both within the acceptable range (Brown, 2015; Hu & Bentler, 1999).

Table S11.6 Fit indices for the confirmatory factor analyses

	χ^2	df	p-value	CFI	RMSEA [90% CI]	SRMR	AIC	BIC	Adjusted BIC
Acceptable fit ¹			> .05	> .95	< .05	< .08	The smaller, the better		
One-factor	6.129	4	.190	.995	.047 [.000, .117]	.020	2,685.907	2,723.136	2,688.278

¹ As per guidelines specified by Hu and Bentler (1999) and Brown (2015).

Note. CFI (Comparative fit index), RMSEA (root mean square error of approximation), SRMR (standardized root mean squared residual), AIC (Akaike information criterion), and BIC (Bayesian information criterion).

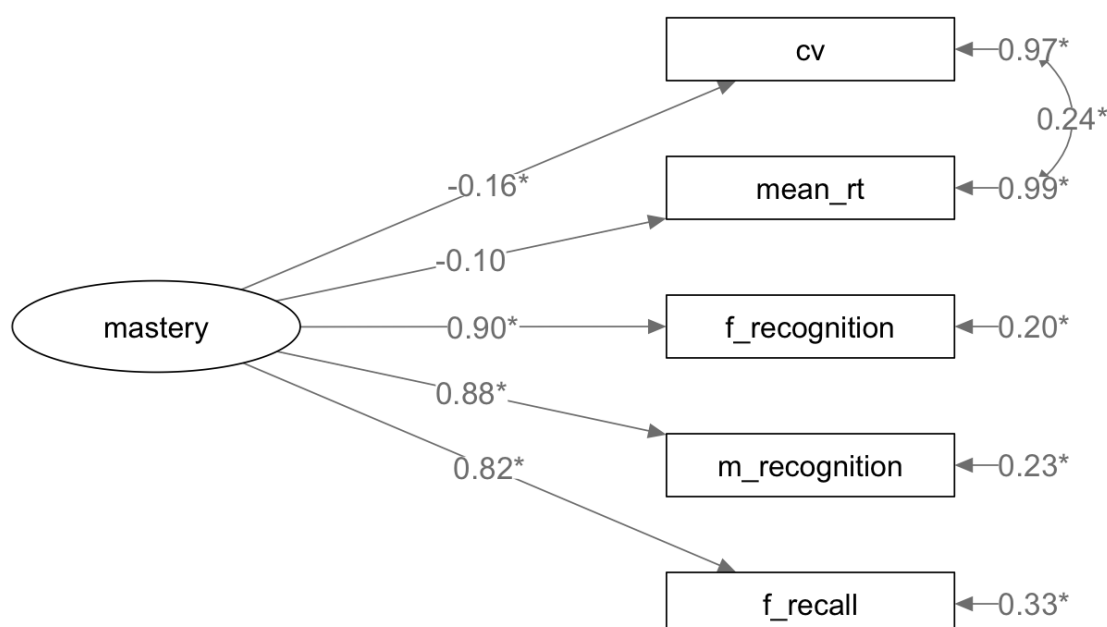


Figure S11.7 Confirmatory factor analysis with standardized factor loadings and error variances. Statistical significance is marked with an asterisk, with exact *p*-values and confidence intervals in Table S11.7. cv = coefficient of variation; mean_rt = mean response time; f_recognition = form recognition; m_recognition = meaning recognition; f_recall = form recall.

Although the one-factor model demonstrated an acceptable fit with the data, it did not achieve convergent validity, as the average variance extracted (.456) fell below the recommended threshold of .5 (Hair et al., 2019). The reliability of the one-factor model was also questionable.

Although the composite reliability coefficient (.626) was slightly above the acceptable threshold (> .6 according to Hair et al., 2019), Cronbach's alpha (.525) fell below this threshold.

The standardized factor loadings between the Lexical Mastery factor and the two processing measures were very weak, negative, and, in the case of mean RT, nonsignificant, suggesting further evidence of model misfit (Figure S11.7 and Table S11.7). This contrasted with the factor loadings for the three knowledge measures, which were very high, positive, and significant ($p < .05$). The high loadings, ranging between .82 and .90, 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.

Table S11.7 Standardized estimates, squared multiple correlations, and z values for the one-factor model

Path	Standardized Estimates [95% CIs]	SE	R^2	z	p
Lexical Mastery =~ Form recall	.816 [.751, .882]	.034	.666	24.292	< .001
Lexical Mastery =~ Meaning recognition	.880 [.831, .928]	.025	.774	35.742	< .001
Lexical Mastery =~ Form recognition	.895 [.843, .947]	.026	.801	33.880	< .001
Lexical Mastery =~ Mean response time	-.104 [-.225, .017]	.062	.011	-1.682	.093
Lexical Mastery =~ Coefficient of variation	-.164 [-.321, -.006]	.081	.027	-2.033	.042
Mean response time ~~ Coefficient of variation	.238 [.052, .425]	.095		2.501	.012

Note. =~ indicates a loading; ~~ indicates a correlation. Boldface indicates statistically significant p -values (i.e., $p < .05$).

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 standardized factor loadings 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.

Research Question 2: 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.

No evidence of multicollinearity was found between predictor variables (variance 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 4 and below into one level in the GCSE listening and reading data to meet the proportional odds assumption for the ordinal models.

Table S11.8 Knowledge and processing as predictors of comprehension

Steps	Predictors	(pseudo) R^2	95% CI	R^2_{adjusted}	ΔR^2	p
<i>DELF listening comprehension</i>						
1	Form recognition, meaning recognition, form recall	.431	[.333, .529]	.423		
2	Form recognition, meaning recognition, form recall, CV	.431	[.333, .528]	.420	.000	1.000
2	Form recognition, meaning recognition, form recall, RT	.431	[.334, .529]	.420	.000	.717
<i>GCSE listening comprehension</i>						
1	Form recognition, meaning recognition, form recall	.496				
2	Form recognition, meaning recognition, form recall, CV	.498			.002	.454
2	Form recognition, meaning recognition, form recall, RT	.499			.003	.309
<i>DELF reading comprehension</i>						
1	Form recognition, meaning recognition, form recall	.531	[.442, .620]	.524		
2	Form recognition, meaning recognition, form recall, CV	.531	[.443, .620]	.523	.000	.660
2	Form recognition, meaning recognition, form recall, RT	.531	[.443, .619]	.522	.000	.913
<i>GCSE reading comprehension</i>						
1	Form recognition, meaning recognition, form recall	.547				
2	Form recognition, meaning recognition, form recall, CV	.548			.001	.940
2	Form recognition, meaning recognition, form recall, RT	.549			.002	.433

Note. To the best of our knowledge, it is not possible to compute 95% CIs around pseudo R^2 (necessitated by the ordinal nature of the GCSE data).

Listening Models

Lexical knowledge as a global construct—combining form recall, meaning recognition, and form recognition—explained 43.09% of the variance in DELF listening (see Table S11.9), with all three predictors reaching significance in the regression model. Of these three predictors, meaning

recognition (15.88%, 95% CI [12.18%, 20.46%]) explained the most variance, followed by form recognition (14.07%, 95% CI [9.39%, 19.75%]), and then form recall (13.15%, 95% CI [8.41%, 19.00%]). Differences in the relative importance of these predictors, however, did not reach significance. The inclusion of CV or mean RT did not significantly increase the variance explained by the model, CV: $F(1, 207) < .001, p = 1.000$; RT: $F(1, 207) = 0.132, p = .717$. See model comparisons in Table S11.8.

Table S11.9 Summary of the Diplôme d’Études en Langue Française (DELF; “Diploma in French Language Studies”) listening regression model (knowledge only)

Predictors	Estimate	DELF listening		RelImp	95% CI	Rank
		95% CI	<i>p</i>			
(Intercept)	−0.00	[−0.10–0.10]	.969			
Form recognition	0.21	[0.02–0.39]	.026	14.07	[9.39, 19.75]	2
Meaning recognition	0.30	[0.12–0.47]	.001	15.88	[12.18, 20.46]	1
Form recall	0.21	[0.05–0.38]	.010	13.15	[8.41, 19.00]	3
Observations	212					
R^2 / R^2 adjusted	.431 / .423					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Table S11.10 Summary of the Diplôme d’Études en Langue Française (DELF; “Diploma in French Language Studies”) listening regression model (knowledge and coefficient of variation)

Predictors	Estimate	DELF listening		RelImp	95% CI	Rank
		95% CI	<i>p</i>			
(Intercept)	−0.00	[−0.11–0.10]	.969			
Form recognition	0.21	[0.02–0.39]	.027	13.93	[9.23, 19.53]	2
Meaning recognition	0.30	[0.12–0.47]	.001	15.73	[11.75, 20.30]	1
Form recall	0.21	[0.05–0.38]	.011	13.08	[8.32, 19.25]	3
Coefficient of variation	−0.00	[−0.11–0.11]	1.000	0.35	[0.06, 2.12]	4
Observations	212					
R^2 / R^2 adjusted	.431 / .420					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Table S11.11 Summary of the Diplôme d'Études en Langue Française (DEL F; “Diploma in French Language Studies”) listening regression model (knowledge and mean response time)

DEL F listening						
<i>Predictors</i>	<i>Estimate</i>	<i>95% CI</i>	<i>p</i>	<i>RelImp</i>	<i>95% CI</i>	<i>Rank</i>
(Intercept)	−0.00	[−0.11–0.10]	.969			
Form recognition	0.21	[0.02–0.39]	.026	14.01	[9.43, 19.77]	2
Meaning recognition	0.30	[0.12–0.47]	.001	15.80	[11.81, 20.52]	1
Form recall	0.21	[0.05–0.37]	.011	13.06	[8.62, 19.00]	3
Mean response time	−0.02	[−0.12–0.08]	.717	0.26	[0.02, 1.96]	4
Observations	212					
R^2 / R^2 adjusted	.431 / .420					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant p -values (i.e., $p < .05$).

A similar pattern emerged for the GCSE listening data. Lexical knowledge explained 49.64% of the variance (see Table S11.12), with form recognition (16.98%) explaining the most, very closely followed by form recall (16.54%) and meaning recognition (16.12%). Like the DEL F listening model, all three predictors reached significance (Table S11.12). Neither CV nor mean RT significantly improved model fit; CV: $\chi^2(1) = 0.560$, $p = .454$; RT: $\chi^2(1) = 1.034$, $p = .309$. See model comparisons in Table S11.8.

Table S11.12 Summary of the General Certificate of Secondary Education (GCSE) listening regression model (knowledge only)

GCSE listening					
<i>Predictors</i>	<i>Odds Ratios</i>	<i>95% CI</i>	<i>p</i>	<i>RelImp</i>	<i>Rank</i>
Form recognition	1.99	[1.23–3.22]	.006	16.98	1
Meaning recognition	1.78	[1.09–3.04]	.029	16.12	3
Form recall	2.03	[1.31–3.17]	.002	16.54	2
Observations	193				
Nagelkerke's R^2	.496				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant p -values (i.e., $p < .05$).

Table S11.13 Summary of the General Certificate of Secondary Education (GCSE) listening regression model (knowledge with coefficient of variation)

GCSE listening					
Predictors	Odds Ratios	CI	<i>p</i>	RelImp	Rank
Form recognition	2.02	[1.25–3.27]	.005	16.97	1
Meaning recognition	1.81	[1.11–3.08]	.024	16.11	3
Form recall	2.00	[1.29–3.12]	.002	16.43	2
Coefficient of variation	1.11	[0.85–1.46]	.456	0.29	4
Observations	193				
Nagelkerke's R^2	.498				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Table S11.14 Summary of the General Certificate of Secondary Education (GCSE) listening regression model (knowledge with RT)

GCSE listening					
<i>Predictors</i>	Odds Ratios	CI	<i>p</i>	RelImp	Rank
Form recognition	1.97	[1.22–3.18]	.006	16.82	1
Meaning recognition	1.78	[1.09–3.06]	.029	16.02	3
Form recall	2.02	[1.31–3.15]	.002	16.41	2
Mean response time	0.85	[0.63–1.16]	.311	0.67	4
Observations	193				
Nagelkerke's R^2	0.499				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Reading Models

Lexical knowledge explained 53.10% of the variance in DELF reading (see Table S11.15), slightly higher than in listening. Reflecting a broadly similar order to listening, meaning recognition (20.49%, 95% CI [16.26%, 25.39%]) accounted for the most variance, closely followed by form recognition (19.30%, 95% CI [14.38%, 24.38%]), and then form recall (13.31%, 95% CI [8.87%, 19.02%]). Only the difference in relative importance between meaning recognition and form recall was significant. Although form recall was not a significant predictor (like form recognition and meaning recognition were), its correlation with DELF reading ($\rho = .604$; see Table S11.5) and dominance weight (13.31%, 95% CI [8.87%, 19.02%]) suggested it explained some variance in performance. As with the listening models, the inclusion of CV or RT did not significantly improve

the model fit, CV: $F(1, 211) = 0.194, p = .660$; RT: $F(1, 211) = 0.012, p = .913$. See model comparisons in Table S11.8.

Table S11.15 Summary of the Diplôme d'Études en Langue Française (DELf; "Diploma in French Language Studies") reading regression model (knowledge only)

DELf reading						
Predictors	Estimate	95% CI	<i>p</i>	RelImp	95% CI	Rank
(Intercept)	−0.00	[−0.09–0.09]	.981			
Form recognition	0.31	[0.15–0.47]	< . .001	19.30	[14.38, 24.38]	2
Meaning recognition	0.36	[0.20–0.52]	< . .001	20.49	[16.26, 25.39]	1
Form recall	0.12	[−0.02–0.27]	.104	13.31	[8.87, 19.02]	3
Observations	216					
R^2 / R^2 adjusted	.531 / .524					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Table S11.16 Summary of the Diplôme d'Études en Langue Française (DELf; "Diploma in French Language Studies") reading regression model (knowledge and coefficient of variation)

DELf reading						
Predictors	Estimate	95% CI	<i>p</i>	RelImp	95% CI	Rank
(Intercept)	−0.00	[−0.09–0.09]	.981			
Form recognition	0.31	[0.14–0.47]	< . .001	19.04	[13.76, 24.83]	2
Meaning recognition	0.35	[0.19–0.51]	< . .001	20.22	[15.46, 25.00]	1
Form recall	0.12	[−0.02–0.27]	.097	13.25	[8.39, 18.63]	3
Coefficient of variation	−0.02	[−0.12–0.07]	.660	0.63	[0.08, 3.04]	4
Observations	216					
R^2 / R^2 adjusted	.531 / .523					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

Table S11.17 Summary of the Diplôme d'Études en Langue Française (DELf; "Diploma in French Language Studies") reading regression model (knowledge and mean response time)

DELf reading						
Predictors	Estimate	95% CI	<i>p</i>	RelImp	95% CI	Rank
(Intercept)	−0.00	[−0.09–0.09]	.982			
Form recognition	0.31	[0.15 – 0.47]	< . .001	19.26	[13.95, 24.95]	2
Meaning recognition	0.36	[0.20 – 0.52]	< . .001	20.46	[16.21, 25.62]	1
Form recall	0.12	[−0.03 – 0.27]	.104	13.28	[8.56, 18.52]	3
Mean response time	0.01	[−0.09 – 0.10]	.913	0.10	[0.03, 1.32]	4
Observations	216					
R^2 / R^2 adjusted	.531 / .522					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

A similar pattern was found for the GCSE reading (Table S11.18), with lexical knowledge accounting for 54.75% of the variance. Form recognition (22.01%) explained the most variance,

followed by meaning recognition (19.75%), and then form recall (12.99%). As with the DELF reading model, form recall did not reach significance, like form recognition and meaning recognition did. However, alternative metrics of predictor importance that are not as sensitive to suppression effects, including Spearman's correlations ($\rho = .595$) and the dominance weight (12.99%), suggested a high degree of interrelatedness between form recall and reading comprehension. Again, CV or RT did not improve the model fit, CV: $\chi^2(1) = 0.006$, $p = .940$; RT: $\chi^2(1) = 0.614$, $p = .433$. See model comparisons in Table S11.8.

Table S11.18 Summary of the General Certificate of Secondary Education (GCSE) reading regression model (knowledge only)

Predictors	GCSE reading				
	Odds Ratios	95% CI	p	RelImp	Rank
Form recognition	2.99	[1.82–4.96]	< .001	22.01	1
Meaning recognition	2.50	[1.44–4.63]	.002	19.75	2
Form recall	1.21	[0.75–1.95]	.428	12.99	3
Observations	193				
Nagelkerke's R^2	.547				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant p -values (i.e., $p < .05$).

Table S11.19 Summary of the General Certificate of Secondary Education (GCSE) reading regression model (knowledge with coefficient of variation)

Predictors	GCSE reading				
	Odds Ratios	CI	p	RelImp	Rank
Form recognition	2.99	[1.83–4.97]	< .001	21.86	1
Meaning recognition	2.51	[1.44–4.63]	.002	19.60	2
Form recall	1.21	[0.75–1.95]	.431	12.92	3
Coefficient of variation	1.01	[0.76–1.35]	.940	0.37	4
Observations	193				
Nagelkerke's R^2	.548				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant p -values (i.e., $p < .05$).

Table S11.20 Summary of the General Certificate of Secondary Education (GCSE) reading regression model (knowledge with RT)

Predictors	GCSE reading			RelImp	Rank
	Odds Ratios	CI	<i>p</i>		
Form recognition	2.97	[1.81–4.95]	< .001	21.88	1
Meaning recognition	2.51	[1.44–4.66]	.002	19.67	2
Form recall	1.21	[0.75–1.95]	.430	12.91	3
Mean response time	0.88	[0.64–1.21]	.435	0.45	4
Observations		193			
Nagelkerke's R^2		.549			

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2). Boldface indicates statistically significant *p*-values (i.e., $p < .05$).

References

- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.
- Hair, J. F., Babin, B. J., Anderson, R. E., & Black, W. C. (2019). *Multivariation data analysis* (8th edition). Cengage.
- 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>
- 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>
- Kenny, D. A. (2024). *Measuring model fit*. <http://davidakenny.net/cm/fit.htm>
- Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th edition). Guilford Press.
- 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>
- 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>

Appendix S12: Alternative Meaning Recall Analyses Using Huibregtse et al.'s (2002) Formula

As acknowledged in the main text, two formulae are in circulation for the index of signal detection (which corrects form-recognition accuracy for guessing behavior). In the Discussion section of the main text, we presented analyses that used the formula reported in Zhang et al. (2020). In this appendix, we present alternative analyses using the following formula, as reported in Huibregtse et al. (2002):

$$1 - \frac{4h(1-f) - 2(h-f)(1+h-f)}{4h(1-f) - (h-f)(1+h-f)}$$

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

As with the alternative analyses presented in Appendix S11, a similar pattern of results emerged, regardless of which formula was used to calculate the index of signal detection.

Descriptive Statistics

Table S12.1 Descriptive statistics and internal consistency reliability for the knowledge measures

	Form Recall	Meaning Recall	Form Recognition (Raw)	Form Recognition (Adjusted ¹)	Meaning Recognition
<i>n</i>	87	87	84	84	87
Mean	66.42%	80.92%	92.09%	80.63%	87.89%
<i>SD</i>	18.92%	17.61%	8.27%	14.03%	13.11%
Median	66.67%	86.67%	93.33%	82.93%	90.00%
95% CI	[62.39%, 70.45%]	[77.17%, 84.67%]	[90.30%, 93.89%]	[77.59%, 83.68%]	[85.10%, 90.69%]
Min	21.67%	26.67%	65.52%	39.80%	26.67%
Max	100.00%	100.00%	100.00%	100.00%	100.00%
Skew	−0.43	−1.11	−1.31	−0.91	−2.06
Kurtosis	−0.53	0.64	1.19	0.37	5.99
Omega	.90	.91	.75	—	.90
Alpha	.89	.89	.72	—	.87

¹ I_{SDT} = index of signal detection.

The internal consistency reliability was generally good-to-excellent for the knowledge measures (Table S12.1). Descriptive statistics demonstrate that learners knew, on average, at least half of the words in each of the knowledge measures and that meaning recognition was the strongest, followed by meaning recall, form recognition (once corrected for false alarm rates), and then form recall, with mostly nonoverlapping CIs between pairs of measures. Note, though, that accuracy rates for meaning recall and form recognition (once corrected for false alarm rates) were very similar. The range in accuracy scores 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 the raw) form recognition scores will be used, as these scores factor in the effect of false alarm rates (i.e., guessing behavior).

Table S12.2 Descriptive statistics and internal consistency reliability for the Diplôme d'Études en Langue Française (DEL F; Diploma in French Language Studies) proficiency measures for a subset of learners who also undertook a meaning-recall test

	Listening	Reading
<i>n</i>	87	87
Mean (%)	51.68%	71.26%
<i>SD</i> (%)	22.40%	24.53%
Median (%)	52.00%	76.00%
95% CI (%)	[46.90%, 56.45%]	[66.04%, 76.49%]
Min (%)	16.00%	12.00%
Max (%)	96.00%	100.00%
Skew	0.19	−0.60
Kurtosis	−1.06	−0.83
Omega	.85	.90
Alpha	.85	.90

Table S12.3 Descriptive statistics for French General Certificate of Secondary Education (GCSE) level

	Percentage of learners achieving each level								Total ¹
	U	3	4	5	6	7	8	9	
Listening	2.70%	4.05%	5.41%	21.62%	4.05%	16.22%	25.68%	20.27%	74
Reading	4.05%	5.41%	8.11%	13.51%	6.76%	6.76%	14.86%	40.54%	74

¹ Not all participants sent an individual breakdown of their GCSE results.

Table S12.4 Spearman’s correlational analyses between the knowledge and proficiency measures

	Form Recall	Meaning Recall	Meaning Recognition	Form Recognition (Raw)	Form Recognition (Adjusted)	DEL F Listening	DEL F Reading	GCSE Listening
Meaning Recall	.768 (< .001)							
Meaning Recognition	.765 (< .001)	.807 (< .001)						
Form Recognition (Raw)	.641 (< .001)	.728 (< .001)	.701 (< .001)					
Form Recognition (Adjusted)	.609 (< .001)	.696 (< .001)	.700 (< .001)	.785 (< .001)				
DEL F Listening	.582 (< .001)	.712 (< .001)	.746 (< .001)	.669 (< .001)	.658 (< .001)			
DEL F Reading	.593 (< .001)	.764 (< .001)	.712 (< .001)	.701 (< .001)	.638 (< .001)	.798 (< .001)		
GCSE Listening	.607 (< .001)	.777 (< .001)	.712 (< .001)	.613 (< .001)	.640 (< .001)	.789 (< .001)	.710 (< .001)	
GCSE Reading	.531 (< .001)	.710 (< .001)	.682 (< .001)	.604 (< .001)	.645 (< .001)	.703 (< .001)	.629 (< .001)	.846 (< .001)

Computed correlation used Spearman method with pairwise deletion.

Note. *p*-values are presented in parentheses in the table above. Boldface indicates statistically significant *p*-values (i.e., $p < .05$). Small: $r > .25$; medium: $r > .40$; large: $r > .60$ (Plonsky & Oswald, 2014). DELF = Diplôme d’Études en Langue Française (“Diploma in French Language Studies”); GCSE = General Certificate of Secondary Education.

Relationship between knowledge measures. Spearman’s correlations (Table S12.4) revealed a high degree of interrelatedness ($r \geq .60$, Plonsky & Oswald, 2014) between the four knowledge measures (form recall, meaning recall, meaning recognition, form recognition), ranging from .609 to .807, but not so strong to raise concerns about multicollinearity ($r \geq .90$; Kline, 2023, p. 56).

Relationship between knowledge and proficiency measures. All four knowledge measures showed highly significant medium-to-large positive correlations with listening measures (Table S12.4), ranging from .582 to .777, and with reading measures, ranging from .531 to .764.

Summary. In sum, we found a very high degree of interrelatedness both within the knowledge measures and between the knowledge and proficiency measures, regardless of modality. The findings from this substudy are thus in line with those reported in the main study.

Relative importance of vocabulary knowledge to listening and reading comprehension

A series of regression models were computed to explore the extent to which the four types of lexical knowledge contributed to L2 listening and reading comprehension. As per the main study, we used dominance weights to ascertain each predictor’s relative importance. For further discussion of this decision, see Analyses in the main body of the manuscript. No evidence of multicollinearity was found between predictor variables ($VIF < 5$ 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 and merged Levels 4 and below into one level to meet the proportional odds assumption for the GCSE ordinal models.

Listening models. Lexical knowledge as a global construct—combining form recall, meaning recall, form recognition, and meaning recognition—explained 52.51% (95% CI [38.91%, 66.11%]) of the variance in the DELF listening model (Table S12.5). Of these four predictors, form recognition explained the most variance, closely followed by meaning recognition, meaning recall, and then form recall.

Table S12.5 Summary of the Diplôme d’Études en Langue Française (DELF; “Diploma in French Language Studies”) listening regression model (knowledge only)

Predictors	Estimate	DELF listening			RelImp	95% CI	Rank
		95% CI	<i>p</i>				
(Intercept)	−0.00	[−0.16–0.15]	.953				
Form recognition	0.32	[0.09–0.55]	.008	15.75	[9.11, 23.22]		1
Meaning recognition	0.27	[−0.07–0.62]	.116	14.17	[9.76, 22.96]		2
Form recall	−0.06	[−0.34–0.23]	.687	8.61	[5.26, 13.71]		4
Meaning recall	0.32	[0.00–0.63]	.050	13.96	[9.31, 21.75]		3
Observations	84						
R^2 / R^2 adjusted	.525 / .501						

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant p -values (i.e., $p < .05$).

A slightly different pattern of results emerged for the GCSE listening model (Table S12.6). Although the four knowledge types explained 57.50% of the variance in the GCSE listening model, a different order of predictor importance emerged. Specifically, we found that meaning recall explained the most variance, followed by meaning recognition, form recognition, and then form recall.

Table S12.6 Summary of the General Certificate of Secondary Education (GCSE) listening regression model (knowledge only)

Predictors	GCSE listening				
	Odds Ratios	95% CI	p	RelImp	Rank
Form recognition	1.84	[0.96–3.60]	.075	11.41	3
Meaning recognition	1.15	[0.46–3.21]	.776	12.63	2
Form recall	0.96	[0.45–2.04]	.917	10.34	4
Meaning recall	5.71	[2.24–15.86]	.001	23.13	1
Observations	73				
Nagelkerke's R^2	.575				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant p -values (i.e., $p < .05$).

Reading models. Lexical knowledge predicted 58.99% (95% CI [46.54%, 71.44%]) of the variance in DELF reading comprehension (Table S12.7) and 55.33% of the variance in GCSE reading comprehension (Table S12.8). In line with the GCSE listening model, we found that meaning recall explained the most variance in DELF and GCSE reading comprehension, followed by meaning recognition, form recognition, and form recall. These findings are thus largely consistent with those reported in Appendix S11.

Table S12.7 Summary of the Diplôme d’Études en Langue Française (DEL F; “Diploma in French Language Studies”) reading regression model (knowledge only)

DEL F reading						
Predictors	Estimate	95% CI	<i>p</i>	RelImp	95% CI	Rank
(Intercept)	−0.01	[−0.16–0.13]	.848			
Form recognition	0.25	[0.03–0.46]	.025	14.03	[7.69, 22.23]	3
Meaning recognition	0.19	[−0.13–0.51]	.241	14.04	[9.26, 22.34]	2
Form recall	−0.10	[−0.36–0.16]	.452	9.91	[6.26, 17.05]	4
Meaning recall	0.55	[0.26–0.84]	< .001	21.01	[13.96, 29.05]	1
Observations	84					
<i>R</i> ² / <i>R</i> ² adjusted	.590 / .569					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., *p* < .05).

Table S12.8 Summary of the General Certificate of Secondary Education (GCSE) reading regression model (knowledge only)

GCSE reading					
Predictors	Odds Ratios	95% CI	<i>p</i>	RelImp	Rank
Form recognition	1.38	[0.68–2.83]	.376	12.81	3
Meaning recognition	3.76	[1.07–14.62]	.054	16.64	2
Form recall	0.79	[0.33–1.85]	.589	8.64	4
Meaning recall	3.36	[1.29–9.48]	.019	17.24	1
Observations	73				
Nagelkerke’s <i>R</i> ²	.553				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight). Boldface indicates statistically significant *p*-values (i.e., *p* < .05).

References

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>

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

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>

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>