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 TBC.

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

Table 1 Components of vocabulary knowledge (Nation, 2013, p. 49)

Form	Spoken	R	What does the word sound like?
		P	How is the word pronounced?
	Written	R	What does the word look like?
		P	How is the word written and spelled?
	Word parts	R	What parts are recognisable in this word?
		P	What word parts are needed to express the meaning?
Meaning	Form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	Concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	Associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	Grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	Collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	Constraints on use	R	Where, when, and how often would we expect to meet this word?
		P	Where, when, and how often can we use this word?

Table 2 Godfroid's (2020, pp. 444–445) expansion of Nation's framework with 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	P	What does the word sound like? Does the learner have a formal-lexical representation (spoken form) of the new word in memory?	Auditory yes/no tests	Auditory lexical decision tasks; auditory priming	Automaticity
		R	How is the word pronounced? How rapidly can the learner produce word forms?	Read aloud exercises	Naming tasks	
	Written	R	What does the word look like? Does the learner have a formal-lexical representation (written form) of the new word in memory?	Proofreading exercises (identifying spelling errors)	Visual lexical decision tasks; form priming; masked repetition priming	
		P	How is the word written and spelled? How rapidly can the learner produce word forms?	Dictation exercises	Written naming (typing) tasks	
	Word parts	R	What parts are recognizable in this word? Does the learner use morphological information to recognize a word?	Word segmentation tasks	Morphological priming	
		P	What word parts are needed to express the meaning? Does the learner put together word parts to produce complex words?	Generating word derivations		
Meaning	Form and meaning	R	What meaning does this word form signal? How rapidly can the meaning of the word be accessed in the lexicon?	Meaning recognition/recall; translation tasks	Lexical decision tasks; semantic categorization tasks; eye tracking tasks	
		P	What word form can be used to express this meaning? 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? How rapidly can the learner access the concept and referents of the word?	Picture-based vocabulary tasks	Visual world eye-tracking paradigm
		P	What items can the concept refer to? How rapidly can the learner produce a word form for a given concept?		Naming tasks
	Association	R	What other words does this make us think of? Has the new word been integrated into an existing semantic network in memory?	Word Associates Test	Semantic priming
		P	What other words could we use instead of this one? How rapidly can the leaner produce word associations?	Synonyms tests	Naming tasks
Use	Grammatical functions	R	In what patterns does the word occur? Is the learner sensitive to ungrammatical uses of words?	Grammaticality judgment tasks	Grammaticality judgment tasks with online measures (self-paced reading, eye tracking, or event-related potentials)
		P	In what patterns must we use this word? Can the learner use the word correctly in real-time conversation?	Guided, untimed writing or speaking tasks (e.g., picture description)	Free, real-time writing 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? How are L2 idioms and L1 idioms in the lexicon related?	Collocation matching tasks	Collocation priming; crosslinguistic priming; eye-tracking
		P	What words or types of words must we use with this one? Do collocations and idioms enjoy a special status in L2 learners' lexicons?	Fill-in-the-black exercises	Naming tasks

Constraints on use	R	Where, when, and how often would we expect to meet this word? Is the learner sensitive to how often words occur and co-occur in the language		Lexical decision tasks; eye tracking
	P	Where, when, and how often can we use this word? Does the learner use words appropriately in context?	Guided, untimed writing or speaking tasks (e.g., picture description)	Free, real-time writing or speaking tasks (e.g., story retelling)

Appendix S2: Information about the GCSE 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 600,000 students) choose to study French, German, and/or Spanish. Prior to taking these exams, students receive approximately 400-to-450 hours of classroom instruction 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 organisations (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 students at Higher tier are expected to achieve Levels 5 to 9. Levels 4 and above are considered a pass, and Levels 5 and above a good pass.

The current subject content (Department for Education, 2015), operational for examinations taken between 2015 and 2025, explicitly outlines the grammar requirements of the qualification but not the vocabulary requirements. In other words, it does not state which lexical items test-takers are expected to know in preparation for their GCSE exams. To assist with planning schemes of learning and textbooks, the awarding organisations publish vocabulary lists in their exam specifications. However, until very recently, due to changes introduced during the pandemic, awarding organisations were required by the government's regulatory body, the Office of Qualifications and Examinations Regulation (Ofqual, 2021), to include words not on the awarding organisations' lists in the exams.

In 2022, the Department for Education in England announced significant 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 word list that awarding organisations must sample from when creating examination papers. A requirement of the revised subject content (Department for Education, 2022) is that at least 85% of the 1,200 and 1,700 lexical items included in the word lists 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).

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 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).

To operationalise this word selection principle, we extracted high-frequency[1] adjectives, nouns, and verbs that were common to the two leading awarding organisations' (AQA and Pearson Edexcel) vocabulary lists created for the purposes of the current GCSE subject content (curriculum), 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 measurements of knowledge and processing. We also ensured that the words appeared on a frequency-informed word list 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 operationalised as within the first 2,000 most frequency words in the French language, according to Lonsdale and Le Bras (2009), as per the guidance example provided in the GCSE Subject Content (2022).

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 3 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 4 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 role of English–French cognates and false friends on accuracy scores (in terms of form recognition, meaning recognition, and meaning recall) 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 and false friends as ANOVA-coded fixed effects, and item and participant ID as random effects. For these analyses only, item scores for meaning 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 5) 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, form recognition, and meaning 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 5 Summary of the regression models investigating the effect of cognates and false friends on accuracy scores

	Form Recall			**Meaning Recognition**			**Form Recognition**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	1.62	0.61 – 4.31	0.335	28.16	12.51 – 63.38	**<0.001**	50.56	21.97 – 116.32	**<0.001**
Cognate Y-N	1.48	0.58 – 3.77	0.407	1.34	0.63 – 2.84	0.450	1.45	0.68 – 3.12	0.337
False Friend Y-N	1.18	0.18 – 7.59	0.862	3.76	0.82 – 17.30	0.089	4.99	1.02 – 24.33	**0.047**
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	0.004 / 0.569			0.015 / 0.521			0.025 / 0.472		

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 6).

Table 6 Summary of the regression model investigating the effect of cognates and false friends on response times

	Response Time		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	808.31	758.73 – 857.89	**<0.001**
Cognate Y-N	-4.03	-49.44 – 41.38	0.862
False Friend Y-N	-52.16	-142.47 – 38.15	0.258
Random Effects			
σ^2	78973.37		
$\tau_{00\ id}$	15762.40		
$\tau_{00\ Target}$	5406.97		
ICC	0.21		
N_{Target}	50		
N_{id}	218		
Observations	9605		
Marginal R^2 / Conditional R^2	0.002 / 0.213		

Appendix S4: Task instructions

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 7 Mark scheme for the form recall task

Score	Criteria
2	- **Target** (correct translation) - **Correct** but different sense (e.g., *vivre* instead of habiter for 'to live') - **Correct or target** word but incorrect / missing accent - **Correct or target** word but incorrect / missing determiner - **Correct or target** but non-default gender (e.g., *voisine* instead of *voisin*)
1	- **Correct or target** word but wrong form (such as plural instead of singular, or past participle instead of infinitive, e.g., *choisi* instead of *choisir*) - **Correct or target** word but irregular plural instead of singular (e.g., yeux instead of œil) - **Correct or target** root but wrong part of speech (e.g., sainement instead of saine) - **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., *à l'étranger* instead of *étranger* (adjective), or *à la mode* instead of *mode, être naître* instead of *naître*)
0-1	- **Correct or target** translation but misspelt (score depends on how badly—give 1 if Levenshtein score >= .7; give 0 if Levenshtein < .7)
0	- Completely non-target (incorrect) - Blank answer - Semantically related but incorrect (e.g., *papier* for 'newspaper') - Multi-word 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). E.g., *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**: e.g., 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. E.g., *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 non-target 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 *cashier* and *cahier*.
- **Multiple answer**: score based on the best answer.

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

In order 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 in-built lm() function in R, with overall accuracy, 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 (Table 8 to Table 10) 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 8 Summary of the regression models investigating the effect of year of data collection on lexical knowledge

	Form Recall			**Form Recognition**			**Meaning Recognition**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	0.00	-0.13 – 0.14	0.962	0.00	-0.13 – 0.13	0.997	0.00	-0.13 – 0.14	0.981
2023 vs. 2022	0.23	-0.03 – 0.50	0.083	0.02	-0.25 – 0.29	0.892	0.12	-0.15 – 0.38	0.393
Observations	218			218			218		
R^2 / R^2 adjusted	0.014 / 0.009			0.000 / -0.005			0.003 / -0.001		

Table 9 Summary of the regression models investigating the effect of year of data collection on lexical processing

	Coefficient of Variation			**Mean Response Time**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	-0.00	-0.14 – 0.13	0.979	-0.00	-0.13 – 0.13	0.995
2023 vs. 2022	-0.13	-0.40 – 0.14	0.344	-0.03	-0.30 – 0.24	0.813
Observations	218			218		
R^2 / R^2 adjusted	0.004 / -0.000			0.000 / -0.004		

Table 10 Summary of the regression models investigating the effect of year of data collection on DELF listening and reading

	DELF listening			**DELF reading**		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	-0.00	-0.14 – 0.14	1.000	0.00	-0.13 – 0.13	0.997
2023 vs. 2022	-0.08	-0.35 – 0.19	0.546	0.02	-0.25 – 0.29	0.862
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 11 Spearman's correlations (with 95% CI) between knowledge, processing, and proficiency measures

Comparison	*rho* [95% CI]	*p* 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-DELF listening	.580 [.483, .663]	< .001
Form-recall-DELF 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-DELF listening	.688 [.610, .753]	< .001
Meaning-recognition-DELF 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)-DELF listening	.645 [.559, .718]	< .001
Form-recognition (raw)-DELF 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)-DELF listening	.666 [.584, .735]	< .001
Form-recognition (adjusted)-DELF 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-DELF listening	-.065 [-.198, .071]	.350
Mean RT-DELF reading	-.066 [-.198, .068]	.331
Mean RT-GCSE listening	-.118 [-.255, .024]	.103
Mean RT-GCSE reading	-.104 [-.242, .037]	.149
CV-DELF listening	-.087 [-.220, .048]	.205
CV-DELF reading	-.118 [-.247, .016]	.085
CV-GCSE listening	-.077 [-.215, .065]	.290
CV-GCSE reading	-.099 [-.237, .043]	.172
DELF listening -DELF reading	.717 [.644, .777]	< .001
DELF listening -GCSE listening	.733 [.660, .793]	< .001
DELF listening -GCSE reading	.633 [.539, .712]	< .001
DELF reading -GCSE listening	.697 [.615, .763]	< .001

DELF reading -GCSE reading	.664 [.576, .736]	< .001
GCSE listening -GCSE reading	.791 [.732, .839]	< .001

Appendix S8: Confirmatory factor analyses

Table 12 Standardised estimates, squared multiple correlations, and z values for the one-factor model

Path	Standardised Estimates [95% CIs]	SE	R^2	z	*p*
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.

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

Table 13 Knowledge and processing as predictors of comprehension

Steps	Predictors	(pseudo) R^2	95% CI	$R^2_{adjusted}$	ΔR^2	*p*
DELF listening comprehension						
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
GCSE listening comprehension						
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
DELF reading comprehension						
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	.550	[.463, .636]	.541	.001	.681
GCSE reading comprehension						
1	Form-recognition, meaning-recognition, form-recall	.567				
2	Form-recognition, meaning-recognition, form-recall, CV	.567			.001	.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% CIs around pseudo R^2 (necessitated by the ordinal nature of the GCSE data).

Table 14 Summary of the DELF listening regression model (knowledge only)

	DELF listening					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.10 – 0.10	0.994			
Form recognition	0.39	0.20 – 0.57	**<0.001**	18.86	[13.25, 25.32]	1
Meaning recognition	0.22	0.05 – 0.39	**0.011**	14.93	[11.31, 19.02]	2
Form recall	0.12	-0.05 – 0.29	0.156	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).

Table 15 Summary of the DELF listening regression model (knowledge and CV)

	DELF listening					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.10 – 0.10	0.990			
Form recognition	0.39	0.20 – 0.57	**<0.001**	18.73	[13.19, 24.94]	1
Meaning recognition	0.22	0.05 – 0.39	**0.013**	14.77	[11.00, 18.75]	2
Form recall	0.12	-0.05 – 0.29	0.155	12.14	[8.19, 16.73]	3
Coefficient of variation	-0.01	-0.11 – 0.10	0.876	0.36	[0.06, 2.40]	4
Observations	212					
R^2 / R^2 adjusted	.460 / .450					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight).

Table 16 Summary of the DELF listening regression model (knowledge and mean RT)

	DELF listening					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.10 – 0.10	0.994			
Form recognition	0.39	0.20 – 0.57	**<0.001**	18.78	[13.29, 24.80]	1
Meaning recognition	0.22	0.05 – 0.39	**0.011**	14.86	[11.29, 19.46]	2
Form recall	0.12	-0.05 – 0.29	0.159	12.14	[8.10, 17.11]	3
Mean response time	-0.01	-0.11 – 0.09	0.875	0.22	[0.03, 1.70]	4
Observations	212					
R^2 / R^2 adjusted	.460 / .450					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight).

Table 17 Summary of the DELF reading regression model (knowledge only)

	DELF reading					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.09 – 0.09	0.985			
Form recognition	0.41	0.24 – 0.58	**<0.001**	22.12	[16.47, 28.26]	1
Meaning recognition	0.32	0.16 – 0.47	**<0.001**	19.87	[15.61, 24.54]	2
Form recall	0.06	-0.09 – 0.21	0.447	12.92	[8.58, 18.16]	3
Observations	216					
R^2 / R^2 adjusted	.549 / .543					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight).

Table 18 Summary of the DELF reading regression model (knowledge and CV)

	DELF reading					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.09 – 0.09	0.985			
Form recognition	0.41	0.24 – 0.58	**<0.001**	21.92	[16.28, 28.18]	1
Meaning recognition	0.31	0.15 – 0.47	**<0.001**	19.56	[15.36, 24.65]	2
Form recall	0.06	-0.09 – 0.21	0.414	12.85	[8.63, 18.12]	3
Coefficient of variation	-0.04	-0.13 – 0.06	0.454	0.69	[0.07, 3.39]	4
Observations	216					
R^2 / R^2 adjusted	.550 / .542					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight)

Table 19 Summary of the DELF reading regression model (knowledge and mean RT)

	DELF reading					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.09 – 0.09	0.988			
Form recognition	0.41	0.24 – 0.58	**<0.001**	22.10	[16.72, 28.34]	1
Meaning recognition	0.32	0.16 – 0.47	**<0.001**	19.84	[15.49, 24.61]	2
Form recall	0.06	-0.09 – 0.21	0.437	12.89	[8.65, 18.28]	3
Mean response time	0.02	-0.07 – 0.11	0.681	0.12	[0.03, 1.25]	4
Observations	216					
R^2 / R^2 adjusted	.549 / .541					

Note. CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight).

Table 20 Summary of the GCSE listening regression model (knowledge only)

	GCSE listening				
Predictors	*Odd Ratios*	*95% CI*	*p*	*RelImp*	*Rank*
Form recognition	3.45	2.04 – 5.93	**<0.001**	22.70	1
Meaning recognition	1.41	0.87 – 2.38	0.173	15.30	3
Form recall	1.62	1.04 – 2.54	**0.036**	15.35	2
Observations	193				
Nagelkerke's R^2	.533				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2).

Table 21 Summary of the GCSE listening regression model (knowledge with CV)

	GCSE listening				
Predictors	Odds Ratios	CI	p	RelImp	Rank
Form recognition	3.48	2.05 – 5.98	**<0.001**	22.66	1
Meaning recognition	1.44	0.89 – 2.41	0.151	15.27	=2
Form recall	1.60	1.03 – 2.51	**0.040**	15.27	=2
Coefficient of variation	1.10	0.84 – 1.45	0.494	0.26	4
Observations		193			
R2 Nagelkerke		0.535			

Table 22 Summary of the GCSE listening regression model (knowledge with RT)

	GCSE listening				
Predictors	*Odds Ratios*	*CI*	*p*	*RelImp*	*Rank*
Form recognition	3.40	2.00 – 5.86	**<0.001**	22.46	1
Meaning recognition	1.42	0.88 – 2.39	0.170	15.20	3
Form recall	1.62	1.04 – 2.54	**0.036**	15.23	2
Mean response time	0.90	0.66 – 1.22	0.487	0.58	4
Observations		193			
R^2 Nagelkerke		0.535			

Table 23 Summary of the GCSE reading regression model (knowledge only)

	GCSE reading				
Predictors	*Odd Ratios*	*95% CI*	*p*	*RelImp*	*Rank*
Form recognition	4.12	2.39 – 7.30	**<0.001**	24.92	1
Meaning recognition	2.19	1.26 – 4.04	**0.009**	19.08	2
Form recall	1.04	0.64 – 1.70	0.860	12.74	3
Observations	193				
Nagelkerke's R^2	0.567				

Table 24 Summary of the GCSE reading regression model (knowledge with CV)

	GCSE reading				
Predictors	*Odds Ratios*	*CI*	*p*	*RelImp*	*Rank*
Form recognition	4.12	2.39 – 7.31	**<0.001**	24.78	1
Meaning recognition	2.18	1.25 – 4.04	**0.010**	18.92	2
Form recall	1.05	0.64 – 1.70	0.856	12.67	3
Coefficient of variation	0.98	0.73 – 1.31	0.882	3.84	4
Observations		193			
R^2 Nagelkerke		0.567			

Table 25 Summary of the GCSE reading regression model (knowledge with RT)

	GCSE reading				
Predictors	*Odds Ratios*	*CI*	*p*	*RelImp*	*Rank*
Form recognition	4.11	2.37 – 7.30	**<0.001**	24.77	1
Meaning recognition	2.20	1.26 – 4.07	**0.009**	19.00	2
Form recall	1.04	0.64 – 1.69	0.865	12.67	3
Mean response time	0.89	0.64 – 1.23	0.473	0.43	4
Observations		193			
R^2 Nagelkerke		0.569			

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 at least 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 randomised 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 26 below presents an overview of the critical words.

Table 26 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) correlations[2], 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 reported as part of the main study.

Descriptive Statistics

Table 27 Descriptive statistics and internal consistency reliability for the knowledge measures

	Form Recall	Meaning Recall	Form Recognition (Raw)	Form Recognition (Adjusted[1])	Meaning Recognition
N	87	87	84	84	87
Mean	66.42%	80.92%	92.09%	72.03%	87.89%
SD	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	.77	-	.90
Alpha	.89	.90	.73	-	.85

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

The internal consistency reliability was generally good-to-excellent for the knowledge measures (Table 27). Descriptive statistics show that, on average, learners knew 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 non-overlapping 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 behaviour).

Table 28 Descriptive statistics and internal consistency reliability for the DELF proficiency measures for a subset of learners who also undertook a meaning recognition test

	Listening	Reading
N	87	87
Mean (%)	51.68%	71.26%
Median (%)	22.40%	24.53%
SD (%)	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 29 Descriptive statistics for French GCSE level

	Percentage of learners achieving each level								
	U	3	4	5	6	7	8	9	Total[1]
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 30 Spearman's correlational analyses between the knowledge and proficiency measures

	Form Recall	Meaning Recall	Form Recognition (Raw)	Form Recognition (Adjusted)	Meaning Recognition	DELF Listening	DELF Reading	GCSE Listening
Meaning Recall	.768 *(<.001)*							
Form Recognition (Raw)	.765 *(<.001)*	.807 *(<.001)*						
Form Recognition (Adjusted)	.641 *(<.001)*	.728 *(<.001)*	.701 *(<.001)*					
Meaning Recognition	.657 *(<.001)*	.741 *(<.001)*	.737 *(<.001)*	.921 *(<.001)*				
DELF Listening	.582 *(<.001)*	.712 *(<.001)*	.746 *(<.001)*	.669 *(<.001)*	.695 *(<.001)*			
DELF 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.

Relationship between knowledge measures. Spearman's correlations (

Table **30**) revealed a high degree of interrelatedness (r ≥. 60, Plonsky & Oswald, 2014) between the four knowledge measures (form recall, meaning recall, meaning recognition, form recognition), ranging from .641 to .807 (

Table **30**), but not so strong to raise concerns about multicollinearity (r ≥. 90; Kline, 2015).

Relationship between knowledge and proficiency measures. All four knowledge measures showed highly significant medium-to-large positive correlations with listening measures (

Table **30**), 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 sub-study 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*.) 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 4[3] 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 comprehension (Table 31). 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 31 Summary of the DELF listening regression model (knowledge only)

	DELF listening					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.15 – 0.15	0.994			
Form recognition	0.39	0.15 – 0.64	**0.002**	18.38	[10.85, 26.79]	1
Meaning recognition	0.28	-0.04 – 0.61	0.088	14.04	[9.49, 23.00]	2
Form recall	-0.09	-0.37 – 0.20	0.548	8.45	[5.26, 12.98]	4
Meaning recall	0.26	-0.06 – 0.57	0.111	13.25	[8.63, 20.10]	3

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

Observations	84
R^2 / R^2 adjusted	.541 / .518

Note: CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight).

A slightly different pattern of results emerged for GCSE listening comprehension (Table 32). Although the four knowledge types explained 58.31% of the variance in the GCSE listening comprehension, 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 32 Summary of the GCSE listening regression model (knowledge only)

	GCSE listening				
Predictors	*Odd Ratios*	*95% CI*	*p*	*RelImp*	*Rank*
Form recognition	2.18	1.07 – 4.54	**0.038**	13.88	2
Meaning recognition	1.18	0.49 – 3.18	0.723	12.40	3
Form recall	0.90	0.42 – 1.93	0.793	10.07	4
Meaning recall	5.12	1.99 – 14.30	**0.002**	21.96	1
Observations	73				
Nagelkerke's R^2	.583				

Note: CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight).

Reading models. Lexical knowledge predicted 60.38% (95% CI [48.20%, 72.55%]) of the variance in DELF reading comprehension (Table 33). In line with the GCSE listening models, we found that meaning recall explained the most variance, followed by form recognition, meaning recognition, and form recall.

Table 33 Summary of the DELF reading regression model (knowledge only)

	DELF reading					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.01	-0.15 – 0.13	0.894			
Form recognition	0.33	0.10 – 0.55	**0.005**	17.20	[10.55, 24.98]	2
Meaning recognition	0.19	-0.12 – 0.49	0.222	13.74	[9.17, 21.33]	3

Form recall	-0.12	-0.38 – 0.14	0.352	9.67	[5.82, 16.31]	4
Meaning recall	0.49	0.20 – 0.79	**0.001**	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).

Although lexical knowledge (as a global construct) predicted 55.58% of the variance in GCSE reading comprehension (Table 34), a slightly different order of importance emerged: Meaning recall predicted 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 34 Summary of the GCSE reading regression model (knowledge only)

		GCSE reading			
Predictors	*Odd Ratios*	*95% CI*	*p*	*RelImp*	*Rank*
Form recognition	1.50	0.72 – 3.19	0.283	13.98	3
Meaning recognition	3.70	1.10 – 13.83	**0.050**	16.42	2
Form recall	0.77	0.32 – 1.81	0.547	8.50	4
Meaning recall	3.20	1.22 – 9.04	**0.025**	16.69	1
Observations	73				
R^2 Nagelkerke	.556				

Note: CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight).

Summary

In sum, the findings reported in this sub-study 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 sub-study, 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.

Appendix S11: Alternative RQ1 and RQ2 analyses 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 behaviour). In the *Results* section of the main text, we presented analyses which used the formula reported in Zhang et al. (2020) and also used in Hui et al. (2025). 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 summarise here. First, form-recognition accuracy (once adjusted for guessing behaviour) was higher by 10.03 percentage points when we used the formula reported in Huibregtse et al. (2002) rather than the formula reported in Zhang et al. (2020) and also used in Hui et al. (2025). 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) and also used in Hui et al. (2025).

Descriptive Statistics

Internal consistency reliability was generally good-to-excellent for the three knowledge measures (Table 35 and Table 36). 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.

Table 35 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 (%)	61.70%	85.63%	89.86%	75.00%
SD (%)	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	.93	.93	.83	-
Alpha	.91	.91	.82	-

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

Table 36 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 37 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 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 task completion may indicate lack of engagement in the task.

Table 38 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.

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 (Table 39). Thus, we refer to CV as a marker of processing stability *and* automaticity, henceforth.

Table 39 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	**.774**								
	(<.001)								
Form-recognition	**.697**	**.753**							
(raw)	**(<.001)**	**(<.001)**							
Form-recognition	**.696**	**.779**	**.745**						
(adjusted)	**(<.001)**	**(<.001)**	**(<.001)**						
Mean response time	-.123	-.131	**-.137**	-.069					
	(.070)	(.053)	**(.044)**	(.311)					
Coefficient of	-.085	-.105	-.032	-.124	**.365**				
variation	(.211)	(.122)	(.640)	(.067)	**(<.001)**				
DELF listening	**.580**	**.688**	**.645**	**.614**	-.065	-.087			
	(<.001)	**(<.001)**	**(<.001)**	**(<.001)**	(.350)	(.205)			
DELF reading	**.604**	**.711**	**.652**	**.671**	-.066	-.118	**.717**		
	(<.001)	**(<.001)**	**(<.001)**	**(<.001)**	(.331)	(.085)	**(<.001)**		
GCSE listening	**.626**	**.705**	**.698**	**.659**	-.118	-.077	**.733**	**.697**	
	(<.001)	**(<.001)**	**(<.001)**	**(<.001)**	(.103)	(.290)	**(<.001)**	**(<.001)**	
GCSE reading	**.595**	**.709**	**.679**	**.727**	-.104	-.099	**.633**	**.664**	**.791**
	(<.001)	**(<.001)**	**(<.001)**	**(<.001)**	(.149)	(.172)	**(<.001)**	**(<.001)**	**(<.001)**

Note. 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).

Relationships Between Knowledge and Processing Measures

Spearman's correlations (Table 39) revealed a high degree of interrelatedness (r ≥. 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 ≥. 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* = . –.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: 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 39). 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.

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 40) 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).

Table 40 Fit indices for the confirmatory factor analyses

	χ2	df	*p* value	χ2 / 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	6.129	4	.190	1.532	.995	.047 [.000, .117]	.020	2,685.907	2,723.136	2,688.278

[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).

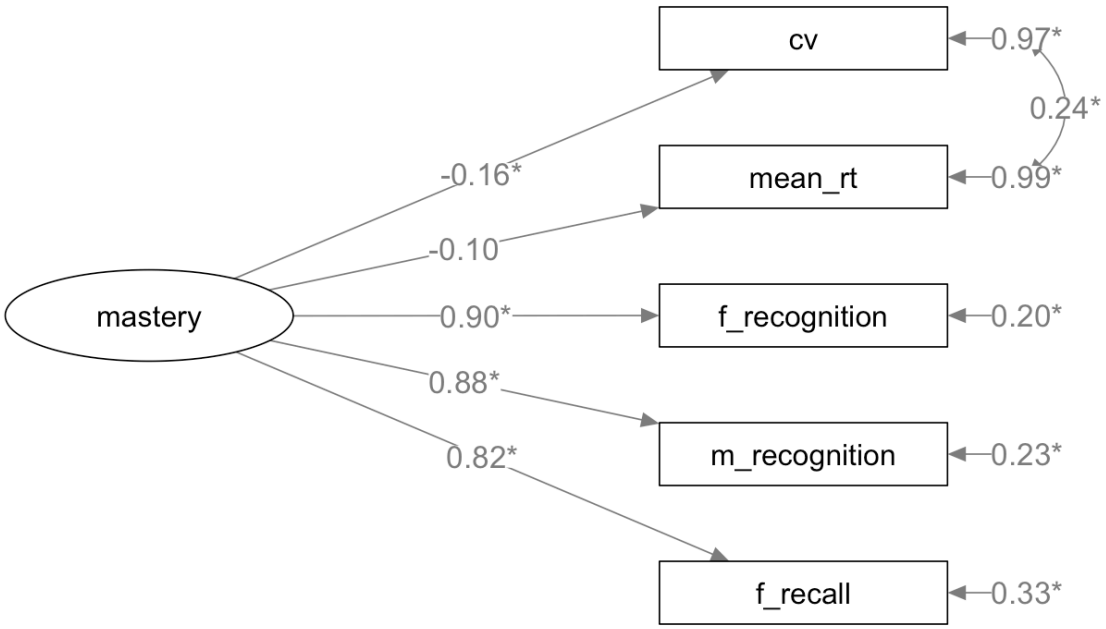

Figure 1 Confirmatory factor analysis with standardised factor loadings and error variances. Statistical significance is marked with an asterisk, with exact *p*-values and confidence intervals in Table 12 of Appendix S8

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 (Awang, 2012). 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 Fornell & Larcker, 1981), Cronbach's alpha (.525) fell below this threshold.

Inspection of the standardised regression coefficients (i.e., paths) between the lexical mastery factor and the two processing measures were very weak, negative, and, in the case of mean RT, non-significant, suggesting further evidence of model misfit (Figure 1 and Table 41). This contrasted with the regression coefficients for the three knowledge measures, which were very high, positive, and significant ($p < .05$). The high loadings ($\beta = .82$ to .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 41 Standardised estimates, squared multiple correlations, and z values for the one-factor model

Path	Standardised 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.

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.

No 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 4[6] and below into one level in the GCSE listening and reading data to meet the proportional odds assumption for the ordinal models.

Table 42 Knowledge and processing as predictors of comprehension

Steps	Predictors	(pseudo) R^2	95% CI	$R^2_{adjusted}$	ΔR^2	*p*
DELF listening comprehension						
1	Form-recognition, meaning-recognition, form-recall	.431	[.333, .529]	.423		
2	Form-recognition, meaning-recognition, form-recall, CV	.431	[.333, .528]	.420	.000	.999
2	Form-recognition, meaning-recognition, form-recall, RT	.431	[.334, .529]	.420	.000	.717
GCSE listening comprehension						
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
DELF reading comprehension						
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, .620]	.522	.000	.913
GCSE reading comprehension						
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 Step 1 in Table 42), with all three predictors reaching significance in the regression model. Of these three predictors (see Table

43), 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 (see model comparisons in Table 13 of Appendix S9; CV: $F(1, 207) = 0.000$, $p = .999$; RT: $F(1, 207) = 0.132$, $p = .719$).

Table 43 Summary of the DELF listening regression model (knowledge only)

		DELF listening				
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.10 – 0.10	0.969			
Form recognition	0.21	0.02 – 0.39	**0.026**	14.07	[9.39, 19.75]	2
Meaning recognition	0.30	0.12 – 0.47	**0.001**	15.88	[12.18, 20.46]	1
Form recall	0.21	0.05 – 0.38	**0.010**	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).

Table 44 Summary of the DELF listening regression model (knowledge and CV)

		DELF listening				
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.11 – 0.10	0.969			
Form recognition	0.21	0.02 – 0.39	**0.027**	13.93	[9.23, 19.53]	2
Meaning recognition	0.30	0.12 – 0.47	**0.001**	15.73	[11.75, 20.30]	1
Form recall	0.21	0.05 – 0.38	**0.011**	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).

Table 45 Summary of the DELF listening regression model (knowledge and mean RT)

		DELF listening				
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.11 – 0.10	0.969			
Form recognition	0.21	0.02 – 0.39	**0.026**	14.01	[9.43, 19.77]	2
Meaning recognition	0.30	0.12 – 0.47	**0.001**	15.80	[11.81, 20.52]	1
Form recall	0.21	0.05 – 0.37	**0.011**	13.06	[8.62, 19.00]	3
Mean response time	-0.02	-0.12 – 0.08	0.717	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).

A similar pattern emerged for the GCSE listening data. Lexical knowledge explained 49.64% of the variance (see Step 1 in Table 42), with form-recognition (16.98%) explaining the most, very closely followed by form-recall (16.53%) and meaning-recognition (16.12%). Like the DELF listening model, all three predictors reached significance (Table 46). Neither CV nor mean RT significantly improved model fit (see model comparisons in Table 42; CV: $\chi^2(1) = 0.560$, $p = .454$; RT: $\chi^2(1) = 0.034$, $p = .309$).

Table 46 Summary of the GCSE listening regression model (knowledge only)

		GCSE listening			
Predictors	*Odd Ratios*	*95% CI*	*p*	*RelImp*	*Rank*
Form recognition	1.99	1.23 – 3.22	**0.006**	16.98	1
Meaning recognition	1.78	1.09 – 3.04	**0.029**	16.12	3
Form recall	2.03	1.31 – 3.17	**0.002**	16.54	2
Observations	193				
Nagelkerke's R^2	.496				

Note. CI = Confidence interval, RelImp = Relative importance (i.e., predictor-specific pseudo R^2).

Table 47 Summary of the GCSE listening regression model (knowledge with CV)

		GCSE listening			
Predictors	Odds Ratios	CI	*p*	RelImp	Rank
Form recognition	2.02	1.25 – 3.27	**0.005**	16.97	1
Meaning recognition	1.81	1.11 – 3.08	**0.024**	16.11	3
Form recall	2.00	1.29 – 3.12	**0.002**	16.43	2
Coefficient of variation	1.11	0.85 – 1.46	0.456	0.29	4
Observations		193			
R2 Nagelkerke		0.498			

Table 48 Summary of the GCSE listening regression model (knowledge with RT)

		GCSE listening			
Predictors	*Odds Ratios*	*CI*	*p*	*RelImp*	*Rank*
Form recognition	1.97	1.22 – 3.18	**0.006**	16.82	1
Meaning recognition	1.78	1.09 – 3.06	**0.029**	16.02	3
Form recall	2.02	1.31 – 3.15	**0.002**	16.41	2
Mean response time	0.85	0.63 – 1.16	0.311	0.67	4
Observations		193			
R^2 Nagelkerke		0.499			

Reading Models

Lexical knowledge explained 53.10% of the variance in DELF reading (Table 49), slightly higher than in listening (see Step 1 in Table 42). 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 (Table 49). Although form-recall was not a significant predictor (like form-recognition and meaning-recognition were), its correlation with DELF reading (*rho* = .604; see

Table **30**) 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 (see model comparisons in Table 42; CV: F(1, 211) = 0.194, p = .660; RT: F(1, 211) = 0.012, p = .913).

Table 49 Summary of the DELF reading regression model (knowledge only)

	DELF reading					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.09 – 0.09	0.981			
Form recognition	0.31	0.15 – 0.47	**<0.001**	19.30	[14.38, 24.38]	2
Meaning recognition	0.36	0.20 – 0.52	**<0.001**	20.49	[16.26, 25.39]	1
Form recall	0.12	-0.02 – 0.27	0.104	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).

Table 50 Summary of the DELF reading regression model (knowledge and CV)

	DELF reading					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.09 – 0.09	0.981			
Form recognition	0.31	0.14 – 0.47	**<0.001**	19.04	[13.76, 24.83]	2
Meaning recognition	0.35	0.19 – 0.51	**<0.001**	20.22	[15.46, 25.00]	1
Form recall	0.12	-0.02 – 0.27	0.097	13.25	[8.39, 18.63]	3
Coefficient of variation	-0.02	-0.12 – 0.07	0.660	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)

Table 51 Summary of the DELF reading regression model (knowledge and mean RT)

	DELF reading					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.09 – 0.09	0.982			
Form recognition	0.31	0.15 – 0.47	**<0.001**	19.26	[13.95, 24.95]	2
Meaning recognition	0.36	0.20 – 0.52	**<0.001**	20.46	[16.21, 25.62]	1
Form recall	0.12	-0.03 – 0.27	0.104	13.28	[8.56, 18.52]	3
Mean response time	0.01	-0.09 – 0.10	0.913	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).

A similar pattern was found for the GCSE reading (Table 52), 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 (Table 52). 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 (see model comparisons in Table 13 of Appendix S9; CV: $\chi^2(1) = 0.006$, $p = .940$; RT: $\chi^2(1) = 0.614$, $p = .433$).

Table 52 Summary of the GCSE reading regression model (knowledge only)

		GCSE reading			
Predictors	*Odd Ratios*	*95% CI*	*p*	*RelImp*	*Rank*
Form recognition	2.99	1.82 – 4.96	**<0.001**	22.01	1
Meaning recognition	2.50	1.44 – 4.63	**0.002**	19.75	2
Form recall	1.21	0.75 – 1.95	0.428	12.99	3
Observations	193				
Nagelkerke's R^2	0.547				

Table 53 Summary of the GCSE reading regression model (knowledge with CV)

	GCSE reading				
Predictors	*Odds Ratios*	*CI*	*p*	*RelImp*	*Rank*
Form recognition	2.99	1.83 – 4.97	**<0.001**	21.86	1
Meaning recognition	2.51	1.44 – 4.63	**0.002**	19.60	2
Form recall	1.21	0.75 – 1.95	0.431	12.92	3
Coefficient of variation	1.01	0.76 – 1.35	0.940	3.76	4
Observations		193			
R^2 Nagelkerke		0.548			

Table 54 Summary of the GCSE reading regression model (knowledge with RT)

	GCSE reading				
Predictors	*Odds Ratios*	*CI*	*p*	*RelImp*	*Rank*
Form recognition	2.97	1.81 – 4.95	**<0.001**	21.88	1
Meaning recognition	2.51	1.44 – 4.66	**0.002**	19.67	2
Form recall	1.21	0.75 – 1.95	0.430	12.91	3
Mean response time	0.88	0.64 – 1.21	0.435	0.45	4
Observations		193			
R^2 Nagelkerke		0.549			

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 behaviour). In the *Discussion* section of the main text, we presented analyses which used the formula reported in Zhang et al. (2020) and Hui et al. (2025). 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.

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 55 Descriptive statistics and internal consistency reliability for the knowledge measures

	Form Recall	Meaning Recall	Form Recognition (Raw)	Form Recognition (Adjusted[1])	Meaning Recognition
N	87	87	84	84	87
Mean	66.42%	80.92%	92.09%	80.63%	87.89%
SD	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	.77	-	.90
Alpha	.89	.90	.73	-	.85

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

The internal consistency reliability was generally good-to-excellent for the knowledge measures (Table 55). Descriptive statistics show that, on average, learners knew 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 non-overlapping 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 behaviour).

Table 56 Descriptive statistics and internal consistency reliability for the DELF proficiency measures for a subset of learners who also undertook a meaning recognition test

	Listening	Reading
N	87	87
Mean (%)	51.68%	71.26%
Median (%)	22.40%	24.53%
SD (%)	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 57 Descriptive statistics for French GCSE level

	Percentage of learners achieving each level								
	U	3	4	5	6	7	8	9	Total[1]
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 58 Spearman's correlational analyses between the knowledge and proficiency measures

	Form Recall	Meaning Recall	Meaning Recognition	Form Recognition (Raw)	Form Recognition (Raw)	DELF Listening	DELF 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)*				
DELF Listening	.582 *(<.001)*	.712 *(<.001)*	.746 *(<.001)*	.669 *(<.001)*	.658 *(<.001)*			
DELF 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.

Relationship between knowledge measures. Spearman's correlations (Table 58) revealed a high degree of interrelatedness (r ≥. 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 ≥. 90; Kline, 2015).

Relationship between knowledge and proficiency measures. All four knowledge measures showed highly significant medium-to-large positive correlations with listening measures (Table 58), 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 sub-study 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*.) Note that little 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 4[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 comprehension (Table 59). Of these four predictors, form recognition explained the most variance, closely followed by meaning recognition, meaning recall, and then form recall.

Table 59 Summary of the DELF listening regression model (knowledge only)

	DELF listening					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.00	-0.16 – 0.15	0.953			
Form recognition	0.32	0.09 – 0.55	**0.008**	15.75	[9.11, 23.22]	1
Meaning recognition	0.27	-0.07 – 0.62	0.116	14.17	[9.76, 22.96]	2
Form recall	-0.06	-0.34 – 0.23	0.687	8.61	[5.26, 13.71]	4
Meaning recall	0.32	0.00 – 0.63	**0.050**	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).

A slightly different pattern of results emerged for GCSE listening comprehension (Table 60). Although the four knowledge types explained 57.50% of the variance in the GCSE listening

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

comprehension, 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 60 Summary of the GCSE listening regression model (knowledge only)

	GCSE listening				
Predictors	*Odd Ratios*	*95% CI*	*p*	*RelImp*	*Rank*
Form recognition	1.84	0.96 – 3.60	0.075	11.41	3
Meaning recognition	1.15	0.46 – 3.21	0.776	12.63	2
Form recall	0.96	0.45 – 2.04	0.917	10.34	4
Meaning recall	5.71	2.24 – 15.86	**0.001**	23.13	1
Observations	73				
Nagelkerke's R^2	.575				

Note: CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight).

Reading models. Lexical knowledge predicted 58.99% (95% CI [46.54%, 71.44%]) of the variance in DELF reading comprehension (Table 61) and 55.33% of the variance in GCSE reading comprehension (Table 62). 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 61 Summary of the DELF reading regression model (knowledge only)

	DELF reading					
Predictors	*Estimate*	*95% CI*	*p*	*RelImp*	*95% CI*	*Rank*
(Intercept)	-0.01	-0.16 – 0.13	0.848			
Form recognition	0.25	0.03 – 0.46	**0.025**	14.03	[7.69, 22.23]	3
Meaning recognition	0.19	-0.13 – 0.51	0.241	14.04	[9.26, 22.34]	2
Form recall	-0.10	-0.36 – 0.16	0.452	9.91	[6.26, 17.05]	4
Meaning recall	0.55	0.26 – 0.84	**<0.001**	21.01	[13.96, 29.05]	1
Observations	84					
R^2 / R^2 adjusted	.590 / .569					

Note: CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight).

Table 62 Summary of the GCSE reading regression model (knowledge only)

		GCSE reading			
Predictors	*Odd Ratios*	*95% CI*	*p*	*RelImp*	*Rank*
Form recognition	1.38	0.68 – 2.83	0.376	12.81	3
Meaning recognition	3.76	1.07 – 14.62	0.054	16.64	2
Form recall	0.79	0.33 – 1.85	0.589	8.64	4
Meaning recall	3.36	1.29 – 9.48	**0.019**	17.24	1
Observations	73				
R^2 Nagelkerke	.553				

Note: CI = Confidence interval, RelImp = Relative importance (i.e., dominance weight).

Appendix S13: References

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

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). *GCSE French, German and Spanish subject content*. https://www.gov.uk/government/publications/gcse-french-german-and-spanish-subject-content

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

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

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

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. (2015). Principles and practices of structural equation modelling. In *Methodology in the social sciences* (4th ed.). Guildford Press.

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

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

Ofqual. (2021). *Ofqual Handbook: General Conditions of Recognition*. https://www.gov.uk/guidance/ofqual-handbook/section-d-general-requirements-for-regulated-qualifications

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

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

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