
Article:
Dudley, Amber orcid.org/0000-0003-2904-9150, Marsden, Emma orcid.org/0000-0003-4086-5765 and Bovolenta, Giulia orcid.org/0000-0003-4139-6446 (2024) A Context-Aligned Two Thousand Test:Towards estimating high-frequency French vocabulary knowledge for beginner-to-low intermediate proficiency adolescent learners in England. Language Testing. pp. 759-791. ISSN 0265-5322

https://doi.org/10.1177/02655322241261415

PREPRINT

Title

A Context-Aligned Two Thousand Test: Towards estimating high-frequency French vocabulary knowledge for beginner-to-low intermediate proficiency adolescent learners in England

11,703 words

Authors

Dr Amber Dudley

Department of Education, University of York

ORCiD: https://orcid.org/0000-0003-2904-9150

Professor Emma Marsden

Department of Education, University of York

ORCiD: https://orcid.org/0000-0003-4086-5765

Dr Giulia Bovolenta

School of Arts, Culture, and Language, Bangor University

ORCiD: https://orcid.org/0000-0003-4139-6446

A Context-Aligned Two Thousand Test: Towards estimating high-frequency French vocabulary knowledge for beginner-to-low intermediate proficiency adolescent learners in England

Abstract

Vocabulary knowledge strongly predicts second-language reading, listening, writing, and speaking. Yet, few tests have been developed to assess vocabulary knowledge in French. The primary aim of this pilot study was to design and initially validate the Context-Aligned Two Thousand Test (CA-TTT), following open research practices. The CA-TTT is a test of written form–meaning recognition of high-frequency vocabulary aimed at beginner-to-low-intermediate learners of French at the end of their fifth year of secondary education. Using an argument-based validation framework, we drew on classical test theory and Rasch modelling, together with correlations with another vocabulary size test and proficiency measures, to assess the CA-TTT's internal and external validity. Overall, the CA-TTT showed high internal and external validity. Our study highlighted the decisive role of the curriculum in determining vocabulary knowledge in instructed, low-exposure contexts. We discuss how this might contribute to under- or over-estimations of vocabulary size, depending on the relations between the test and curriculum content. Further research using the tool is openly invited, particularly with lower-proficiency learners in this context. Following further validation, the test could serve as a tool for assessing high-frequency vocabulary knowledge at beginner-to-low-intermediate levels, with due attention paid to alignment with curriculum content.

Introduction

Vocabulary knowledge strongly predicts second language (L2) proficiency in listening (In'nami et al., 2022), reading (Jeon & Yamashita, 2022), writing (Kojima et al., 2022), and

speaking (Jeon et al., 2022). This is not surprising given that learners need to know at least 95% of the words in any given written or spoken text in English to fully understand it (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). Knowledge of high-frequency words can thus help learners to reach this level of coverage. For instance, the first 2,000 most frequent words in English have been found to cover at least 82% of written language and 89% of spoken language (Dang & Webb, 2014; Webb & Nation, 2017; Webb & Rodgers, 2009). As such, there is a need for learners to know and be tested on their knowledge of high-frequency words (Webb et al., 2017). Indeed, many education systems prioritise high-frequency words in their curricula. Highly relevant to the current study are the Department for Education's (2022) recently announced reforms to the General Certificate in Secondary Education (GCSE; a national high-stakes external examination taken almost exclusively by 16-year-olds in schools) curriculum for French, German, and Spanish in England. The revised curriculum stipulates that the words used in each exam must be sampled from a compulsory wordlist where at least 85% of the items are high-frequency (defined as being in the most frequent 2,000 words). For further information about GCSEs, curriculum reforms, and other jurisdictions where frequency has informed vocabulary selection, see Appendix S1.

Although many vocabulary tests are available in English (Meara & Milton, 2003; Nation, 1990; Nation & Beglar, 2007; Schmitt et al., 2001), very few exist for languages other than English. This is problematic for several reasons. First, it makes it difficult to know whether the reported relations between English vocabulary knowledge and L2 proficiency similarly hold for learners of languages other than English. Second, teachers and materials and test developers currently do not have a reliable understanding of vocabulary knowledge among these learners that could help when selecting or creating appropriate materials (Nation & Beglar, 2007; Schmitt et al., 2001; Stoeckel & Bennett, 2015). Although some such tests do exist (for French, see Batista & Horst, 2016; Milton, 2006; Peters et al., 2019), they provide

limited coverage of the first 2,000 most frequent words and/or are not designed specifically with beginner-to-low intermediate adolescent learners in instructed contexts in mind. This gap is of particular concern given that current test development and validation theory advocates against a one-size-fits-all approach (Chapelle, 2012; Kane, 2006; Read, 2000). Likewise, there have been calls for more rigorous validation of existing *and* new tests of vocabulary knowledge and, in particular, a better specification of tests' purpose(s) and the type of learners and educational contexts that tests have been developed for (Schmitt et al., 2020).

To address these gaps, we piloted a new written test of form–meaning recognition of high-frequency vocabulary for beginner-to-low intermediate learners of French: the Context-Aligned Two Thousand Test (CA-TTT). This article describes the rationale and process behind the CA-TTT's development and presents results from a pilot study designed to initially validate this test with 222 16-year-old English-speaking learners of French. Using an argument-based validation framework, we drew on classical test theory and Rasch modelling, together with correlations with another vocabulary size test and proficiency measures, to assess preliminary evidence for the CA-TTT's internal and external validity.

Literature

This section reviews the existing measures of form–meaning knowledge in English and French that motivated the CA-TTT's development and then outlines the argument-based validation framework used in this study.

Tests of vocabulary knowledge in English

When developing measures of vocabulary knowledge, most researchers have adopted a frequency-driven approach to item selection. Perhaps the most well-known tests of form(–meaning) recognition in English are the X-Lex (Meara & Milton, 2003), Y-Lex (Meara &

Miralpeix, 2006), Vocabulary Levels Test (Nation, 1990; Schmitt et al., 2001), and Vocabulary Size Test (Nation & Beglar, 2007).

X-Lex (Meara & Milton, 2003) is a yes–no (self-report) test of form recognition knowledge. In each of the three versions, participants are presented with 120 words: 20 words from the first five 1,000-word frequency bands and 20 pseudowords and are told that not all words are real and must tick the words they know or can use. For every real word ticked, 50 points are awarded and for every pseudoword ticked, 250 points are deducted to account for false alarms. X-Lex suits low-proficiency learners due to its low cognitive demands. X-Lex, however, tests form recognition, not form–meaning recognition, and can thus only give a partial indication of vocabulary knowledge. Y-Lex (Meara & Miralpeix, 2006) adopts an identical format, but where X-Lex focuses on the 5,000 most frequent words, Y-Lex tests vocabulary in the 6,000-10,000 word frequency range and may therefore be better suited to more advanced learners.

The original VLT estimates learners' written receptive knowledge of the form–meaning links of words in four frequency bands (2,000, 3,000, 5,000, and 10,000) and an academic vocabulary level, whereas the updated VLT (Webb et al., 2017) focuses on the first five bands (1,000, 2,000, 3,000, 4,000, and 5,000). In both versions, each band includes 30 items consisting of five 6-noun clusters, three 6-verb clusters, and two 6-adjective clusters. Within each cluster, learners must select which of the six words matches one of three definitions. The VST (Nation & Beglar, 2007), on the other hand, estimates learners' written receptive vocabulary size and contains 140 items sampled from the 14,000 most frequently occurring words in English with 10 items per frequency band. The VST has since been expanded to include the 20,000 most frequently occurring words with five items per frequency band (Coxhead et al., 2015). Within each band, participants must select which of the four definitions matches the target word presented within a sentence. The form–meaning recognition format,

however, has been criticised for several reasons, including its potential to overestimate vocabulary knowledge and lower internal reliability relative to more open-ended formats such as meaning recall (McLean et al., 2020; Stewart et al., 2023).

Most of this test development research has focused on English, with few measures of form(–meaning) recognition being available for learners of other languages, including French. Further research is thus needed in languages other than English.

Tests of vocabulary knowledge in French

The available measures for French include X-Lex (Meara & Milton, 2003; Milton, 2009), the Test de la Taille du Vocabulaire (TTV; Batista & Horst, 2016), and the VocabLab test (Peters et al., 2019).

The French X-Lex (Meara & Milton, 2003; Milton, 2009) is similar to the English version. The three versions (forms) tests knowledge of the 5,000 most frequent words in French sampled from Baudot's (1992) frequency list. Several studies (David, 2008; Milton, 2006, 2015) have used X-Lex to estimate vocabulary size among GCSE French learners. Milton (2006) found that these learners (n = 49) knew approximately 852 words (standard deviation [SD] = 440, range: 0-1,800). In a follow-up study, Milton (2015) reported similar findings: 775 words (n = 18, SD = 341, range: 350-1,250). David (2008) found even lower sizes: 564 (n = 26, SD = 352, range: 0-1,650), although the discrepancy is likely due to learners being tested at the beginning of the school year in David's study and at the end in Milton's.

The VocabLab test (Peters et al., 2019), of which one version (form) exists, assesses form–meaning recognition among Dutch-speaking learners of French. The test samples 30 words from each four frequency bands (2,000, 3,000, 4,000, and 5,000) based on the Lonsdale and Le Bras (2009) frequency list. The 2,000 band is broader than the others and includes words from both the 1,000 and 2,000 bands. In this test, participants select a word's meaning from

four options, but unlike the original VLT, words are presented in isolation (rather than a sentence) and an 'I don't know' option is included to reduce guessing. The use of an 'I don't know' option is not without criticism due to individual differences in how likely participants are to select it (Zhang, 2013). Nevertheless, weaker correlations between proficiency and vocabulary knowledge have been reported when the option is included relative to when it is not (Stoeckel et al., 2016). The VocabLab test, however, is not a measure of vocabulary *size* as it does not contain a dedicated 1,000 band. As such, accuracy rates per frequency band, not estimated vocabulary sizes, are reported.

The TTV (Batista & Horst, 2016), of which one version (form) exists, adopts a similar format to the VLT and tests 120 words, with 30 words from each four frequency bands (2,000, 3,000, 5,000, 10,000) based on Lonsdale and Le Bras' (2009) frequency list. The items in the 10,000 band, however, are from the Baudot (1992) frequency list, as the Lonsdale and Le Bras list only contains the 5,000 most frequent words. Unlike the VocabLab test, the TTV does not include any items from the 1,000 band.

Limitations of existing tests

The VocabLab, the TTV, and X-Lex are not without their limitations. Although the former two include more items from each frequency band (i.e., 30 instead of 20 in the latter), following recent recommendations (Gyllstad et al., 2021; Schmitt et al., 2020), these items were *randomly* sampled from each band without consideration for the vocabulary learners might encounter in the classroom. This design feature may be inherent to the very purpose of a size test. However, it can cause problems if these tests are administered in specific populations. For instance, although X-Lex has been used to test GCSE learners' vocabulary knowledge, only 27% of the 100 test items appeared in at least one or more of the vocabulary lists created for these learners (Assessment and Qualifications Alliance [AQA], 2016; Edexcel, 2018).

Critically, by the time that 16-year-olds in England take their GCSE exams, they will have received approximately 400-to-450 hours of classroom exposure to French, with very little (if any) exposure outside of the classroom. These learners' lexicons are thus largely restricted to the classroom input, which is typically composed of the vocabulary featured in the GCSE curriculum lists, the textbooks written using those lists, and the GCSE exam papers. Moreover, much of this vocabulary is likely to be mid-to-low frequency: Marsden, Dudley, and Hawkes (2023), for instance, reported that of the 1,322 flemmas on AQA's (the leading awarding organisation in England) current GCSE French wordlist, only 48% were high-frequency. Thus, any test of vocabulary knowledge that *randomly* samples 20 or even 30 words from each band is unlikely to provide a valid or useful measure of these students' vocabulary knowledge.

Such an argument echoes recent calls to examine the role of factors beyond frequency alone in predicting word difficulty (Hashimoto, 2021). For instance, He and Godfroid (2019) gathered usefulness and difficulty ratings from 76 experienced teachers of L2 English and found that frequency correlated only moderately with perceived usefulness and difficulty. Likewise, Robles-García et al. (2023) observed that 29 teachers' judgments of what words their students were most likely to know had a stronger relationship with students' vocabulary test scores than frequency. These findings point to the influence of classroom instruction on students' vocabulary knowledge.

Another limitation of existing tests is that they "lack the needed precision to estimate the number of words that a learner knows [and] to determine mastery of specific word bands" (Stoeckel et al., 2021, p.198). One way to address these limitations in light of the above discussion, at least with classroom learners with limited L2 exposure, may be to develop measures that factor in word frequency *and* the language featured in the curriculum.

A commonly cited advantage of vocabulary size tests is that they can assess the knowledge of learners from a wide range of proficiencies. However, their design often means that they provide more useful information about the vocabulary knowledge of intermediate-to-advanced learners and/or learners who have ample exposure to language outside the classroom than for the beginner-to-low intermediate proficiency level and limited exposure that characterise GCSE learners.

First, the tests provide limited coverage of the 2,000 most frequent words, despite their high importance for comprehension. For instance, neither the VocabLab test nor the TTV has a dedicated 1,000 band: The 2,000 band in the VocabLab test sampled 30 words from the 0-2,000 range (i.e., approximately 15 words in each 1,000 band), and the 2,000 band in the TTV only sampled 30 items from the 1,000-2,000 band. There is thus a need to develop a test that focuses solely on assessing high-frequency vocabulary knowledge, particularly in instructed contexts such as ours where a compulsory list of high-frequency vocabulary has recently been introduced for those starting to study GCSE French in 2024.

Second, both the VocabLab and TTV tests provide definitions for target words in the L2. Thus, each item tests knowledge of the target word and the words in the multiple-choice options (i.e., definitions). As Elgort (2013) argues, vocabulary size estimates using bilingual tests—where the target word is presented in the L2 and the multiple-choice options in the first language (L1)—are likely to be larger and more accurate especially among intermediate (and, even more so, beginner) learners. It is therefore not surprising that many bilingual versions of the English VST have been developed (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011), although not yet for French.

Finally, there are no data about how the TTV performs with a population comparable to GCSE learners. Although the TTV reliably distinguished between proficiency levels, it was only validated with university-level students (Batista & Horst, 2016). In contrast, the VocabLab

test was validated with different age groups, including secondary school students. These groups generally had lower scores and displayed more variability than other groups. Given the low-proficiency characterising their secondary school group, Peters et al. (2019) argued that it may be worthwhile to develop a test that focuses specifically on the 1,000 most frequent words. This has also been suggested by other researchers, including Webb and Sasao (2013).

Considering these issues, we developed the Context-Aligned Two Thousand Test (CA-TTT) to assess knowledge of curriculum-relevant words from the 2,000 most frequent French words among beginner-to-low intermediate adolescent learners. In our preliminary validation of the CA-TTT, we chose to focus on learners who had recently completed their GCSE French exams. In doing so, we acknowledge that further research will be needed to validate the test with more diverse sets of beginner-to-low intermediate learners.

GCSEs are high-stakes national exams taken by approximately 600,000 16-year-olds in at least one (academic) subject every year. For most students, these subjects include Maths, English, and Science, together with five optional subjects. Approximately 20% of each annual cohort choose to study French as an optional subject. These numbers, however, have decreased dramatically in the past two decades from 331,089 in 2003 to 130,901 in 2023 (Joint Council for Qualifications, 2003, 2023). Concerns have thus been raised about a shortage of language skills in the UK and its impact on the country's ability to compete internationally (Ayres-Bennett et al., 2022). With this in mind, we focused predominately on testing high-achieving learners (i.e., those obtaining Level 7 or above in GCSE French) in the current study, as these individuals are the most likely to pursue further language study and, in turn, help to address the current shortage in language skills.

Given the size of this population, our limited understanding of these learners' vocabulary knowledge, and their impact on the UK's language skills shortage, the CA-TTT's intended uses were to provide: (a) a test instrument for *researchers* to explore the extent to

which high-frequency, curriculum-relevant vocabulary knowledge correlates with existing measures of vocabulary knowledge and L2 proficiency, (b) a valuable datapoint to inform *policy-makers*' decisions regarding language learning, teaching, and testing, and, eventually following further validation work, (c) an achievement test for *teachers* to identify gaps in students' curriculum-specific knowledge as they approach their high-stakes exams.

Argument-based Validation

The process of validation involves collating evidence to support and explain the interpretation of a test's scores for its intended purpose (Purpura et al., 2015). A unitary view of test validation, as proposed by Messick (1989) and Kane (2006, 2013), has become highly influential in language testing. Kane's argument-based framework (2006, 2013) is based upon an interpretive validity argument whereby test designers must explicitly state their claims about test score interpretation and use and then provide a series of inferences about the test—that is, justifications supported by logical and/or empirical evidence. Adaptations of this framework has been successfully applied in many domains of L2 research, including Bokander and Bylund's (2020) validation of the LLAMA language aptitude test. In a similar vein, our study adopts a logical framework of argument-based validation that is described in Table 1.

Table 1. Proposed framework for the validation of the CA-TTT.

Domain description inference	
Sample of test items	Do observations of test performance reflect relevant knowledge and skills used in situations representative of those in the target language use (TLU) domain? *(test design, stimuli selection)*
Generalisation inference	
Universe score	Is the test representative of all possible samples of universe items? *(internal consistency, model fit)*
Scoring inference	
Observed score	Is the test made up of good quality items? *(item fit statistics, discrimination index, item difficulty)*

Explanation inference	
Construct interpretation	Are the tasks reasonable given theories or constructions? Is there an explanation for misfitting items? *(content analysis, scale unidimensionality)* Is the test associated in a way that is consistent with previous L2 vocabulary research? *(interaction with other components of vocabulary knowledge such as form recognition)*
Extrapolation inference	
Target score	Do test scores reflect success in various areas or levels of L2 learning? *(correlation with L2 behaviour in specific language tests)* Do test scores reflect L2 learning success? *(correlation with general L2 skills)*

Note: In italics are the types of evidence that might be considered in the current validation process.
Source: Adapted (to reflect the purposes of the current study) from Table 1 in Bokander and Bylund (2020, p.18), a study that drew on works by Kane (2006, 2013) and Purpura et al. (2015).

The current study

The purpose of the current pilot study was to design and initially validate a test of context-aligned high-frequency vocabulary knowledge for beginner-to-low intermediate school-aged learners of French. In doing so, we also sought to explore the extent to which different approaches to sampling of test items can affect vocabulary size estimates in instructed, low-exposure contexts.

To achieve this, we set out to assess four test-internal links (domain description, scoring, generalisation, explanation) and one test-external (higher-order) link (extrapolation) in the chain of inferences, using the validation framework presented in Table 1. Specifically, at the level of domain description inference, we compared the level of overlap between the test items and the vocabulary used in the target language use (TLU) domain (i.e., the curriculum followed by the participants in this study) to determine whether observations of test performance revealed relevant knowledge in situations representative of those in the TLU domain. Second, we examined the generalisation inference, using internal consistency

measures, and the scoring inference, using Rasch modelling to assess whether the test was made up of items of appropriate difficulty. Then, at the level of explanation inference, we conducted item content analyses to explain any misfitting items and correlated CA-TTT performance with X-Lex estimates, a measure of form recognition. Finally, we investigated the extrapolation inference by examining the extent to which CA-TTT estimates correlated with performance in high-stakes and standardised testing. Our research questions (RQs), generated from the validation framework proposed in Table 1, were as follows:

RQ1 (domain description inference): To what extent do CA-TTT items reflect the vocabulary used in the TLU (that is, the curriculum underlying their high-stakes GCSE exams)?

RQ2 (generalisation inference): To what extent do CA-TTT scores exhibit internal consistency?

RQ3 (scoring inference): To what extent is the difficulty of the CA-TTT appropriate to the beginner-to-low intermediate proficiency level of the GCSE learners tested in the current study after approximately 400-to-450 hours of exposure to classroom instruction?

RQ4 (explanation inference): What is the strength of association between CA-TTT performance and performance on X-Lex, a test of form recognition?

RQ5 (extrapolation): What is the strength of association between CA-TTT scores and performance in high-stakes GCSE exams and in standardised tests of receptive proficiency?

In the past decade, open research practices have been gaining traction in the language sciences (Liu et al., 2022; Marsden & Morgan-Short, 2023), with an increasing number of materials, data, and analysis code being made Findable, Accessible, Interoperable, and

Reusable (FAIR; GO FAIR, n.d.). Exemplifying these FAIR principles, this article shares the materials, data, and analysis code used to initially validate the CA-TTT via our Open Science Framework repository (https://osf.io/k4y7p/) and on Instruments and Data for Research in Language Studies (IRIS, n.d.). Data cleaning and analysis were conducted using the freely available statistical software, R, to ensure that the analysis pipeline is reusable.

Method

Participants

Participants included two cohorts of 16-year-old learners of French (113 in 2022 and 109 in 2023) who had recently (within the previous one-to-six weeks) finished their GCSE exams, after approximately 400-to-450 hours of instruction in French and very little (if any) exposure outside the classroom. For more information about learners' language background and minimal out-of-school exposure, see Appendix S2. On average, participants reported learning French from 9.68 years of age (95% CI [9.31, 10.06], SD = 2.83, range: 1-15). All participants (of which 26% reported English as an additional language) had completed their secondary education in English and were from 89 state-funded secondary schools from across England. Participants were recruited via their school and told that participation was optional and that they would receive £25 or £35 in Amazon vouchers for completing two or three sessions, respectively. Ethics approval for the study was obtained from the University of York.

Instruments and Procedures

The Context-Aligned Two Thousand Test

Test items

Given the low-proficiency of our target population and the importance of high-frequency words for comprehension, the CA-TTT focuses solely on the 2,000 most frequently occurring lemmas

from the Lonsdale and Le Bras (2009) frequency list. The lemma includes the base form (e.g., dance) and its inflections (e.g., dances, dancing, danced). Acknowledging the ongoing debate surrounding lexical units (Kremmel, 2021; Webb, 2021), the lemma was selected as many learners do not possess the relevant knowledge to comprehend the derivational forms of known headwords (Brown et al., 2022).

We removed the 3,000 to 10,000 bands from Batista and Horst's (2016) TTV and created a new 1,000 band while still sampling from the same frequency list (Lonsdale & le Bras, 2009) as the original TTV. Thus, the CA-TTT contains two bands: 1-1,000 and 1,001-2,000. Each band in the CA-TTT is twice as large as each band in the TTV, with 60 target items per band split across 20 clusters and 120 items in total. (We supplemented the 30 target items in the existing 2,000 band with an additional 30 items.) The number of items per 1,000-word frequency band was based on Gyllstad et al.'s (2021) recommendation that researchers use *at least* 30 items because test score inferences become more representative of actual knowledge as the number of items increase. The CA-TTT maintains a 3:5:2 ratio between verbs, nouns, and adjectives across clusters, respectively, broadly mirroring the part of speech distributions of the Lonsdale and Le Bras' (2009) word frequency list. When sampling new items, we included as many words as possible from the awarding organisations' (AQA, 2016; Edexcel, 2018) GCSE vocabulary lists (of 1,058 and 1,811 lemmas, respectively) to approximate the vocabulary used in the classroom. (For more information about how these lists were developed; see Appendix S1, Marsden et al. [2023]; Finlayson et al. [2024]). To make the test more sensitive to partial knowledge and less demanding for beginner-to low intermediate learners, word definitions were presented in English (Elgort, 2013).

Test format

English definitions were presented in clusters of three alongside a drop-down menu from which participants could choose one of six French words from the same part of speech (Figure 1).

Participants were told to “choose the French word closest to the word or phrase on the right”. The words in the drop-down menu were identical for each definition in the cluster, but their order was randomised across definitions within each cluster.

Answers were scored on a binary scale (1 for correct word–definition matches; 0 for incorrect matches). Estimates of high-frequency vocabulary size were inferred by multiplying the decimal percentage of correctly answered items from each frequency band by the number of words (1,000) in each band (Batista & Horst, 2016).

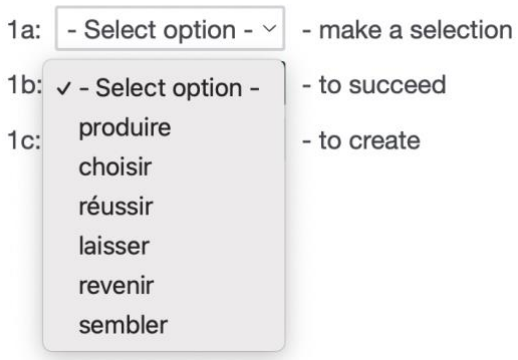

Figure 1. Sample CA-TTT items from one cluster

Form–meaning recognition format was selected over recall primarily due to the ease and simplicity with which it could be administered and scored. However, in selecting a test of form–meaning recognition as opposed to recall, we acknowledge its limitations, including the reportedly lower internal reliability of this format relative to more open-ended formats (McLean et al., 2020) and its potential to overestimate vocabulary size (Gyllstad et al., 2021; Stoeckel et al., 2021). At the same time, we note that these findings almost exclusively pertain to research conducted among adult highly-educated L2 learners of English. Until further research is undertaken, we argue that form–meaning recognition remains a valid and thus appropriate measure of vocabulary knowledge.

X-Lex Test

We also administered the French X-Lex Vocabulary Test (the first version ["Test 1"] as reported by Milton, 2009; available via FLLOC, n.d.). This test had a very low overlap with the CA-TTT: Of the 40 items from the 0-2,000 range in X-Lex—the range relevant to the CA-TTT—none were used as target words in the CA-TTT and only two (*ville* 'town' and *peser* 'to weigh') were used as distractors.

In the 2022 iteration of the current study, the test consisted of 100 real words and 20 pseudowords randomised across participants. In the 2023 iteration, we included an additional 20 pseudowords to align with recommendations in the field (Pellicer-Sánchez & Schmitt, 2012). These additional pseudowords, however, did not appear to influence vocabulary size estimates (see Appendix S3 for the full analyses). Participants saw the following instructions: "Please look at these words. Some of these words are real French words and some are invented but are made to look like real words. Please tick the words that you know or can use". Although the presence of "or" may result in ambiguity, the original instructions were maintained.

X-Lex was scored following the procedure described by Milton (2006, 2015). The number of 'Yes' responses to real words was multiplied by 50 to give a maximum raw size estimate of 5,000. The number of 'Yes' responses to pseudowords was then multiplied by 250 for the 2022 dataset (and 125 for the 2023 dataset; given the higher number of pseudowords, this maintained parity with the calculation across iterations) and subtracted from the raw score to account for false alarms. Unlike previous studies (David, 2008; Milton, 2006, 2015), participants were not excluded if they ticked five or more pseudowords given the potential for such data trimming to over-estimate vocabulary size.

Figure 2. Sample X-Lex items. Pseudowords are italicised for illustration purposes only.

DELF proficiency test

Participants completed the listening and reading sections of the Common European Framework of Reference for Languages A2 Junior version of the *Diplôme d'études en langue française* (DELF; France Éducation International, n.d.), a French proficiency test that participants were not familiar with. In this test, participants read and listened to short passages and answered multiple-choice comprehension questions. The DELF was selected due to it meeting all 17 of the Association of Language Testers in Europe's quality standards as well as the ease with which it can be administered and scored. We were specifically interested in the listening and reading components given our focus on the receptive form–meaning link and its strong relationship with listening and reading (Zhang & Zhang, 2022).

Procedure

The tests were administered online through the survey platform, Qualtrics (n.d.), between June and August in 2022 and in 2023 as part of a larger study on the components of French language proficiency among GCSE students. This larger study consisted of two 90-minute sessions and one further optional session. The CA-TTT and X-Lex were completed in the first session, and the DELF sub-tests in the second. In August, we asked participants to self-report their GCSE results, including their overall and skill-specific (listening, reading, writing, speaking) levels (graded as 1–9), by providing a photo of their official results statement.

When designing the study, we were faced with the challenge of testing this population at the height of their knowledge—that is, in the summer holidays following their GCSE exams. Although participants were not monitored when completing the tasks, to mitigate risk of cheating, participants were told at the beginning of each session that they would not receive compensation for their involvement in the study if they consulted external sources (e.g., the Internet, friends, or family). Additional measures, including disabling the copy and paste function within Qualtrics and forcing the browser into full screen mode, were implemented.

Results

Score Overview

Table 2 and

Table 3 present raw accuracy scores and estimated vocabulary sizes, respectively. Shapiro-Wilk tests revealed significant deviations from normality both in the 1,000 ($W = .829$, $p < .001$) and 2,000 band ($W = .865$, $p < .001$). Inspection of histograms (Figure 3) and skewness coefficients further showed that scores in both frequency bands were negatively skewed, thus suggesting that the test was easy for most participants.

Table 2. CA-TTT raw and percentage accuracy scores.

Frequency Band (*k*=60)	%	Mean	SD	95% CIs	Min	Max	Skew	Kurt
1,000	86.96%	52.18	7.88	[51.14, 53.21]	17.00	60.00	-1.72	6.45
2,000	75.43%	45.26	10.07	[43.93, 46.58]	0.00	59.00	-1.68	6.79
Total	81.19%	48.72	9.67	[47.82, 49.62]	0.00	60.00	-1.63	6.76

Table 3. CA-TTT estimated vocabulary size.

Frequency Band	Mean estimate	SD	95% CIs	Min	Max	Skew	Kurt
1,000	870	131	[853, 887]	283	1000	-1.72	6.45
2,000	754	168	[732, 776]	0	983	-1.68	6.79
Total	1,624	285	[1586, 1662]	466	1983	-1.51	5.34

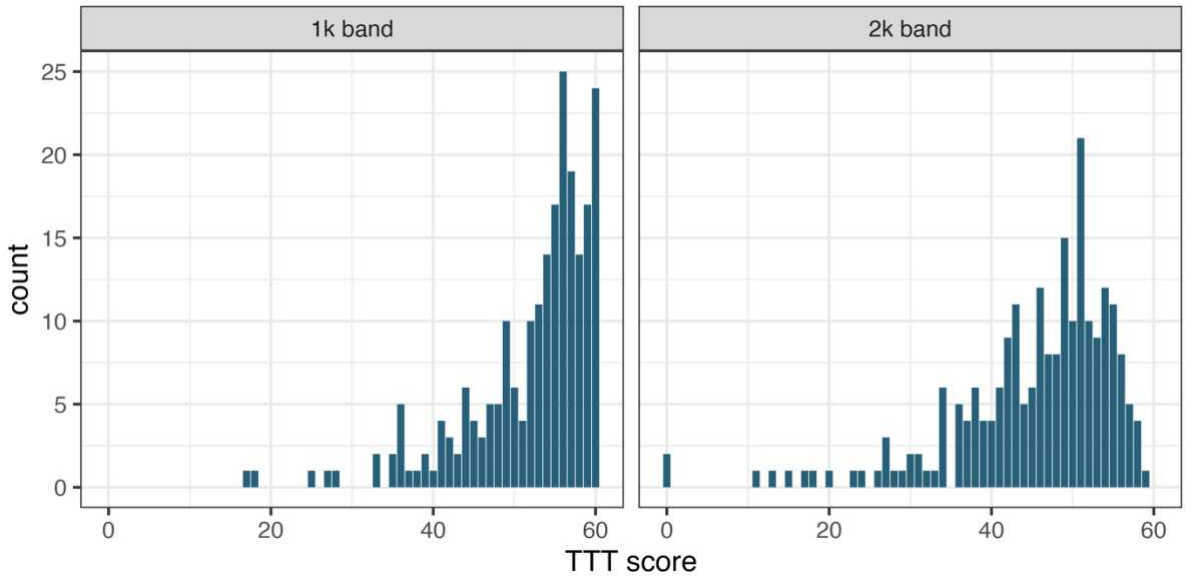

Figure 3. Distribution of CA-TTT raw scores.

Domain Description Inference.

To investigate whether CA-TTT test items were representative of the vocabulary used in the TLU domain (RQ1), we calculated the level of overlap between CA-TTT items and vocabulary on the GCSE curriculum list. As expected, given our approach to sampling, the level of overlap was very high: On average, 79.42% (SD = 2.61%, 95% CI [79.07%, 79.77%]) of CA-TTT items also appeared on the curriculum list specific to the participant, their awarding organisation (AQA or Edexcel), and entry tier (foundation or higher).

We then explored the extent to which this overlap determined CA-TTT performance. Non-overlapping confidence intervals (Table 4) suggested that mean accuracy scores were consistently higher for words that were on the relevant list than those that were not.

Table 4. Accuracy scores for words on and off the relevant curriculum list.

	Mean	SD	95% CI	Min	Max
On the list (mean k^1=*95*)	84.61%	14.01%	[82.74%, 86.47%]	28.09%	100.00%
Off the list (mean k^1=*25*)	68.83%	17.01%	[66.57%, 71.09%]	0.00%	100.00%

[1] We report mean *k* as the number of words on the curriculum list was specific to the awarding organisation and entry tier for which participants were entered.

To further examine the predictive role of the curriculum, we then adjusted the initial CA-TTT vocabulary size estimates by multiplying (a) the mean accuracy decimal percentage for words *on* the list in the CA-TTT by the number of high-frequency words *on* the relevant curriculum list and (b) the mean accuracy decimal percentage for words *off* the list in the CA-TTT by the number of high-frequency words *off* the curriculum list, and then adding the two estimates together (see Table 5 for raw and adjusted estimates). For example, if a participant scored 80% for the words on the curriculum list and 50% for the words off the curriculum list, the corresponding decimal percentages would be multiplied by the number of high-frequency words (out of the total 2,000) on (649) and off (1,351) the curriculum list respectively: (0.80×649)+(0.50×1,351)=1194.70. This process accounted for words being more likely to be known if they were on the curriculum list to provide a more objective measure of known high-frequency words. These calculations resulted in a significant decrease in vocabulary size estimates, as demonstrated by non-overlapping confidence intervals (Table 5).

Table 5. Raw and adjusted vocabulary size estimates from the CA-TTT

	Mean	SD	95% CI	Min	Max	Skew	Kurt
Unadjusted	1,627	285	[1,589, 1,664]	467	1,983	-1.53	2.40
Adjusted	1,489	299	[1,449, 1,529]	303	1,993	-1.38	2.51

Generalisation Inference

To explore the question of generalisation and, in particular, the internal consistency of the test, we computed categorical omega for the overall test and each frequency band in two steps, following Flora (2020).[1] We first fitted a one-factor confirmatory factor analysis using the *lavaan* package (Rosseel, 2023) to test the unidimensionality assumption for omega and Rasch models (i.e., to see whether all items loaded onto a single factor). A one-factor model (Table 6) was a good fit for the overall test and both frequency bands, suggesting that the items measured the same construct and thus met the unidimensionality assumption. We then obtained omega estimates using the *reliability()* function from the *semTools* package (Jorgensen et al., 2022). These estimates indicated good reliability (Nunnally & Bernstein, 1994): .92 for the 1,000 band, .94 for the 2,000 band, and .96 for the two bands combined.

Table 6. Fit indices for the one-factor confirmatory factor analysis fitted to CA-TTT items.

Band	Tucker-Lewis Index	Comparative Fit Index	Root Mean Square Error of Approximation
Accepted cut-off criteria (Hu & Bentler, 1999)	> .95	> .95	< .06
1,000 (*k*=60)	.949	.950	.017
2,000 (*k*=60)	.970	.971	.017
Overall (*k*=120)	.951	.950	.014

Since 26% of our sample reported having a first language (L1) other than English, we explored the effect of language background on CA-TTT performance. Overlapping confidence intervals around mean accuracy percentages for learners with L1 English and an L1 other than English suggested no significant difference (see Appendix S4).

Scoring Inference

To address whether the CA-TTT is made up of items of appropriate difficulty (RQ3; the scoring inference), we compared Rasch model estimates from two packages: *eRm* (Mair et al., 2021), a conditional maximum likelihood estimation package, and *TAM* (Robitzsch et al., 2022), a joint maximum likelihood estimation package, following recent guidance to conduct both (Linacre, 2021; Nicklin & Vitta, 2022). Given the largely negligible differences in estimates, we present the *eRm* models here and the corresponding *TAM* models in Appendix S5.

To test the local independence assumption (Baghaei, 2008), a pre-requisite for Rasch modelling, we inspected correlations between test item residuals. Residuals were not significantly correlated (overall test: mean p = .48 [SD = .29]; 1,000 band: mean p = .50 [*SD* = .29]; 2,000 band: mean p = .48 [*SD* = .29]), suggesting that our data met the local independence assumption.

To visualise how difficult specific items were for individual participants, person and item values were plotted together on the same logit scale in individual Wright maps for each frequency band (see Figure 4 and Figure 5). Items were plotted on the *y*-axis and the latent dimension (item difficulty/person ability) on the *x*-axis. The histogram at the top shows the distribution of person abilities. A participant placed at the same point on the scale as an item has a 50% probability of getting that item right. If a participant is placed higher on the scale than the item is, then the chance of the participant getting the item right is above 50%. In

contrast, if a participant is placed lower on the scale than the item is, their chance of getting the item right is below 50%. For the item and person parameters, see Appendix S6.

Although item difficulties were evenly distributed, the test appeared to be very easy for the vast majority of the sample: In most cases, the item means (i.e., 0 on the *x*-axis) were below many of the participants' chances of getting that item right. Both in the band-specific and overall Rasch models, mean person ability was higher than maximum item difficulty, especially in the 1,000 band. As expected, the 2,000 band was more challenging than the 1,000 band, with a greater overlap between item difficulty and person ability distributions (i.e., a smaller distance between mean item difficulty and mean person ability relative to the 1,000 band) due to a higher proportion of challenging items.

To examine how reliably the test could distinguish between different abilities, we calculated person separation reliability. The value for both the 2,000 band (.88) and the overall test (.93) indicated two or more separate levels of performance in the data. In contrast, the value for the 1,000 band (.80) was on the threshold between low and acceptable separation reliability (Aryadoust et al., 2021), indicating a lack of discrimination between high and low ability participants due to the relative ease of the frequency band.

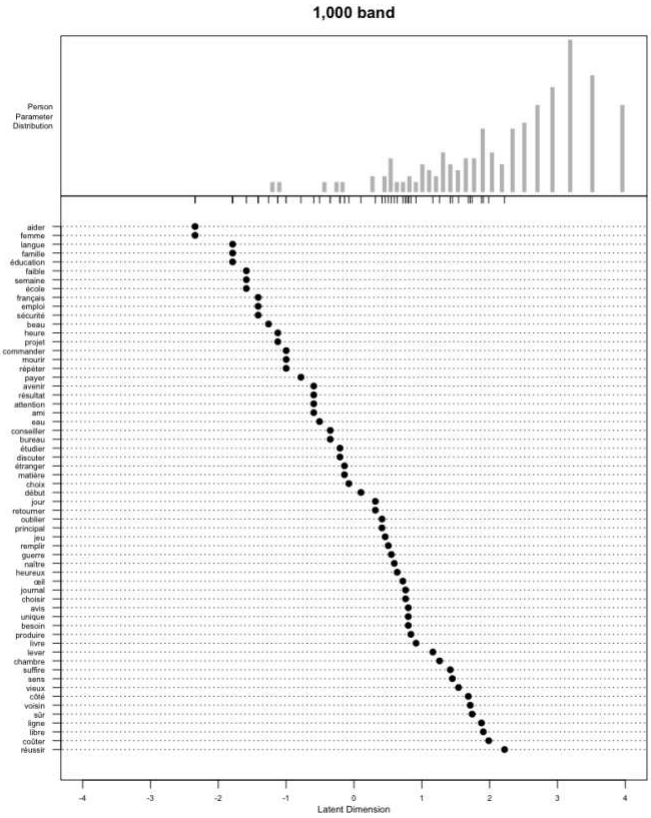

Figure 4. Wright map for items in the 1,000 band.

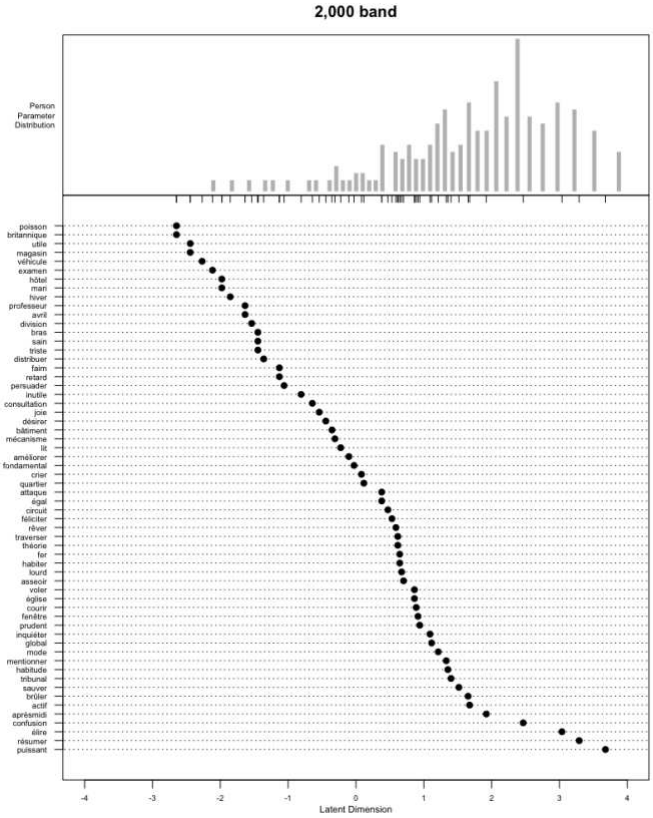

Figure 5. Wright map for items in the 2,000 band.

Explanation Inference

To examine the explanation inference (RQ4), we conducted content analyses to identify any potentially misfitting items and then explored whether the test was associated in a manner consistent with previous L2 vocabulary research, by examining correlations between CA-TTT and X-Lex scores.

Content Analysis: Facility and discriminations indices

Infit mean square values (see Appendix S6) for CA-TTT items were all within the optimal range (i.e., between 0.5 and 1.5; Linacre, 2002; Wright & Linacre, 1994) both in the frequency-band models and the overall test model. There was, however, greater variation in outfit mean square values with both underfitting and overfitting items. According to Wright and Linacre (1994), values below 0.5 (underfitting) and between 1.5 and 2 (overfitting) are unproductive (but not degrading) for the construction of measurement.

The overall Rasch model identified three items with outfit mean-square statistics above 2: *éducation* 'education' (with the correct response being 'learning'), *femme* 'woman' ('adult female'), and *puissant* 'powerful' ('which has great power'). Both *éducation* and *femme* were easy items, with facility indices of .98 and .99 and estimated logit (difficulty) values of -1.79 and -2.83, respectively. Participants who answered incorrectly (n = 5 for *éducation* and n = 3 for *femme*) included those who were within the bottom 10th percentile of performers or who scored 90% or more. In contrast, *puissant* was a difficult item, with a facility index of .17 and an estimated logit value of 4.19. The band-specific Rasch models revealed a similar pattern of results. Although *femme* was not identified as a misfitting item, *semaine* 'week' ('seven days') was. *Semaine* was an easy item, with a facility index of .97 and an estimated logit value of -1.59. Again, the five participants who answered incorrectly included those who were within the bottom 10th percentile of performers or who scored 90% or more. Since outfit mean-square statistics are sensitive to mistakes by more-proficient learners (i.e., outlier gaps between item

difficulty and person ability; Linacre, 2002), this may explain the poor fit exhibited by these four items.

Correlation with another vocabulary test: X-Lex

To further address the explanation inference, we analysed correlations between (unadjusted) CA-TTT (Table 3) and X-Lex estimates (Table 7) for each frequency band. (For full X-Lex scores, see Appendix S7). Given that X-Lex and the CA-TTT sample from different frequency bands (X-Lex: 1,000 to 5,000; CA-TTT: 1,000 and 2,000), we only compared performance on the 1,000 and 2,000 bands, not *overall* scores from the two tests. To obtain comparable estimates, we divided the overall pseudoword penalty by five (the number of bands in X-Lex) to get a 'by-band' pseudoword penalty estimate and subtracted this value from raw scores for the 1,000 and 2,000 bands to calculate adjusted X-Lex scores (henceforth, vocabulary size estimates).

Table 7. X-Lex vocabulary size estimates (penalty adjusted scores).

Frequency Band (k = 20)	Mean estimate	SD	95% CI	Min	Max	Skew	Kurt
1,000	437	181	[413, 461]	-50	850	-0.25	-0.15
2,000	273	186	[249, 298]	-100	750	0.28	-0.54
Total	711	340	[666, 756]	-150	1,600	0.07	-0.14

The Wilcoxon test for paired samples showed that mean vocabulary size estimates significantly differed between the CA-TTT and X-Lex ($V = 24{,}753$, $p < .001$). However, strong positive correlations (

Figure 6) were found between the CA-TTT and X-Lex for the 1,000 band (*rho* = .67, 95% CI [.59, .74], $p < .001$) and the 2,000 band (*rho* = .69, 95% CI [.61, .75], $p < .001$). (Spearman's *rho* was used due to both estimates being non-normally distributed.)

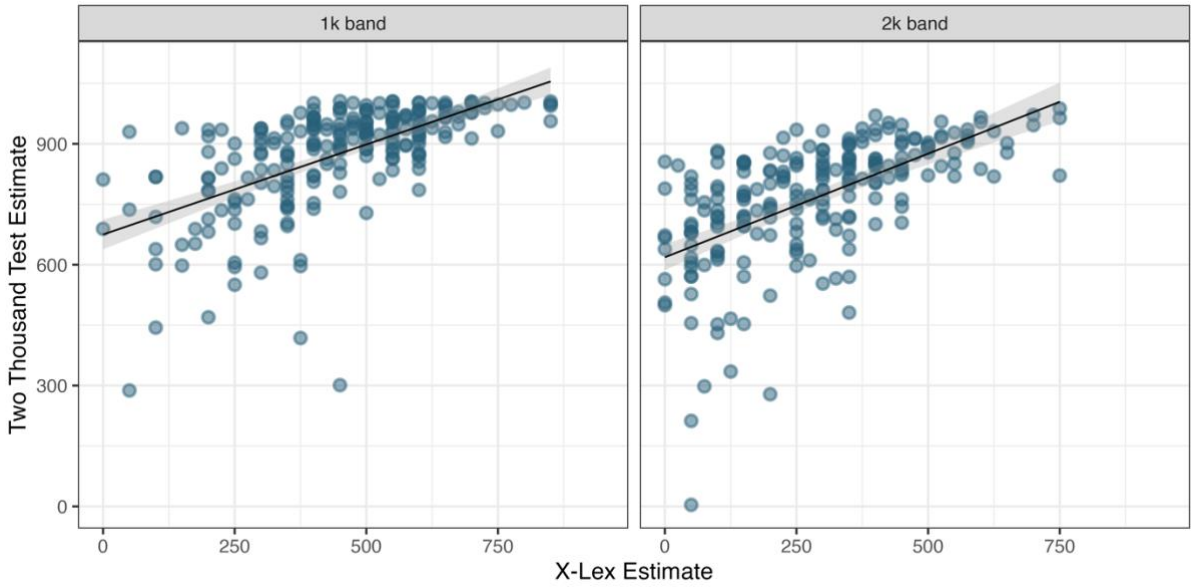

Figure 6. Scatterplots showing associations between the CA-TTT and X-Lex scores. Darker dots represent a higher number of observations.

Differences in mean estimates between the two measures were very likely due to differences in test items. On average, of the 40 X-Lex items in the 1,000 and 2,000 frequency bands, only 25.81% (SD = 2.91%, 95% CI [25.42%, 26.20%]) appeared on the GCSE curriculum list, a level of overlap significantly lower than that observed for the CA-TTT (M = 79.42%, SD = 2.61%, 95% CI [79.07%, 79.77%]). To explore the role of the curriculum further, we compared mean accuracy percentages for words on and off the list. Non-overlapping confidence intervals around the mean suggested that participants were more likely to know a word in X-Lex if it appeared on (M = 53.57%, SD = 22.14%, 95% CI [50.63%, 56.52%]) than off (M = 46.63%, SD = 20.00%, 95% CI [43.97%, 49.29%]) the curriculum list.

Extrapolation Inference

To examine the extrapolation inference (RQ5), we analysed the relationships between (unadjusted) CA-TTT estimates and proficiency measures from both high-stakes (GCSE scores) and standardised testing (DELF scores), given that vocabulary knowledge strongly predicts L2 proficiency (see Introduction).

Table 8. GCSE French levels.

	Percentage achieving each level										
	U	1	2	3	4	5	6	7	8	9	*n*
Reading	2%	0%	0%	4%	4%	14%	5%	8%	18%	44%	195
Listening	3%	0%	0%	3%	6%	17%	4%	21%	19%	28%	195
Speaking	0%	0%	0%	2%	5%	14%	11%	8%	15%	44%	195
Writing	1%	0%	0%	1%	6%	15%	10%	12%	13%	41%	195
Overall	0%	0%	<1%	<1%	5%	17%	9%	12%	22%	35%	220

GCSE levels. Of the 222 participants, 220 (99%) self-reported their overall level and 195 (88%) reported a skill breakdown (Table 8Table 8). Because GCSE data were ordinal and CA-TTT data non-normally distributed, Spearman's correlations were calculated, using the *cor.ci()* function from the *psych* package (Revelle, 2024). CA-TTT estimates had strong positive correlations (>.60; Plonsky, 2015) with overall and skill-specific level (Table 9). That is, students with larger CA-TTT estimates were more likely to obtain higher GCSE grades in each skill than those with smaller CA-TTT estimates.

Table 9. Spearman's correlations between CA-TTT estimates and GCSE levels.

	rho	*95% CI*
Overall	.77*	[.69, .83]
Listening	.72*	[.64, .79]
Reading	.73*	[.65, .78]
Speaking	.62*	[.52, .70]
Writing	.66*	[.55, .74]

* $p < .001$

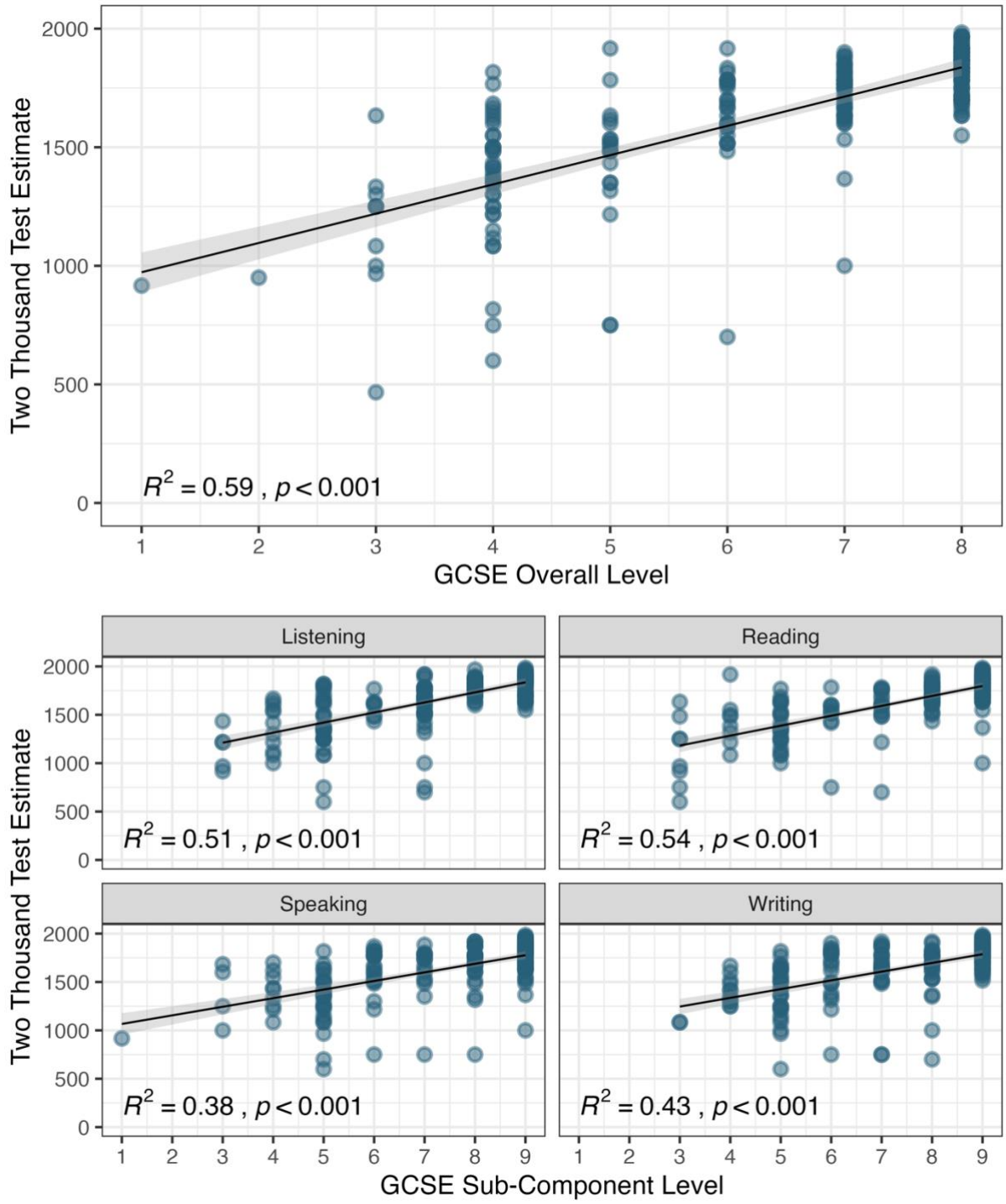

Figure 7. Scatterplots showing associations between CA-TTT estimates and GCSE French performance.

Table 10. DELF (receptive) raw scores (n = 222)

Test (total)	Mean	SD	95% CI	Min	Max	Skew	Kurt	α	ω
Listening (/25)	13.05	6.03	[12.26, 13.85]	0	25	0.14	-1.00	.85	.85
Reading (/25)	17.74	5.91	[16.96, 18.52]	2	25	-0.63	-0.68	.90	.90
Overall (/50)	30.79	10.92	[29.35, 32.24]	8	50	-0.20	-1.03	.93	.93

DELF scores. Finally, we explored the relations between DELF scores (Table 10) and CA-TTT estimates (Figure 8). Because CA-TTT estimates were non-normally distributed, Spearman's correlations were computed, using the *cor.ci()* function from the *psych* package (Revelle, 2024). CA-TTT estimates demonstrated a strong positive correlation (>.60; (Plonsky, 2015) with overall DELF scores and skill-specific scores (Table 11). That is, students who scored highly on the CA-TTT also scored highly on the DELF measures.

Table 11. Spearman's correlations between CA-TTT estimates and DELF scores.

	rho	*95% CI*
Overall	.77*	[.68, .83]
Listening	.68*	[.59, .75]
Reading	.75*	[.66, .82]

* $p < .001$

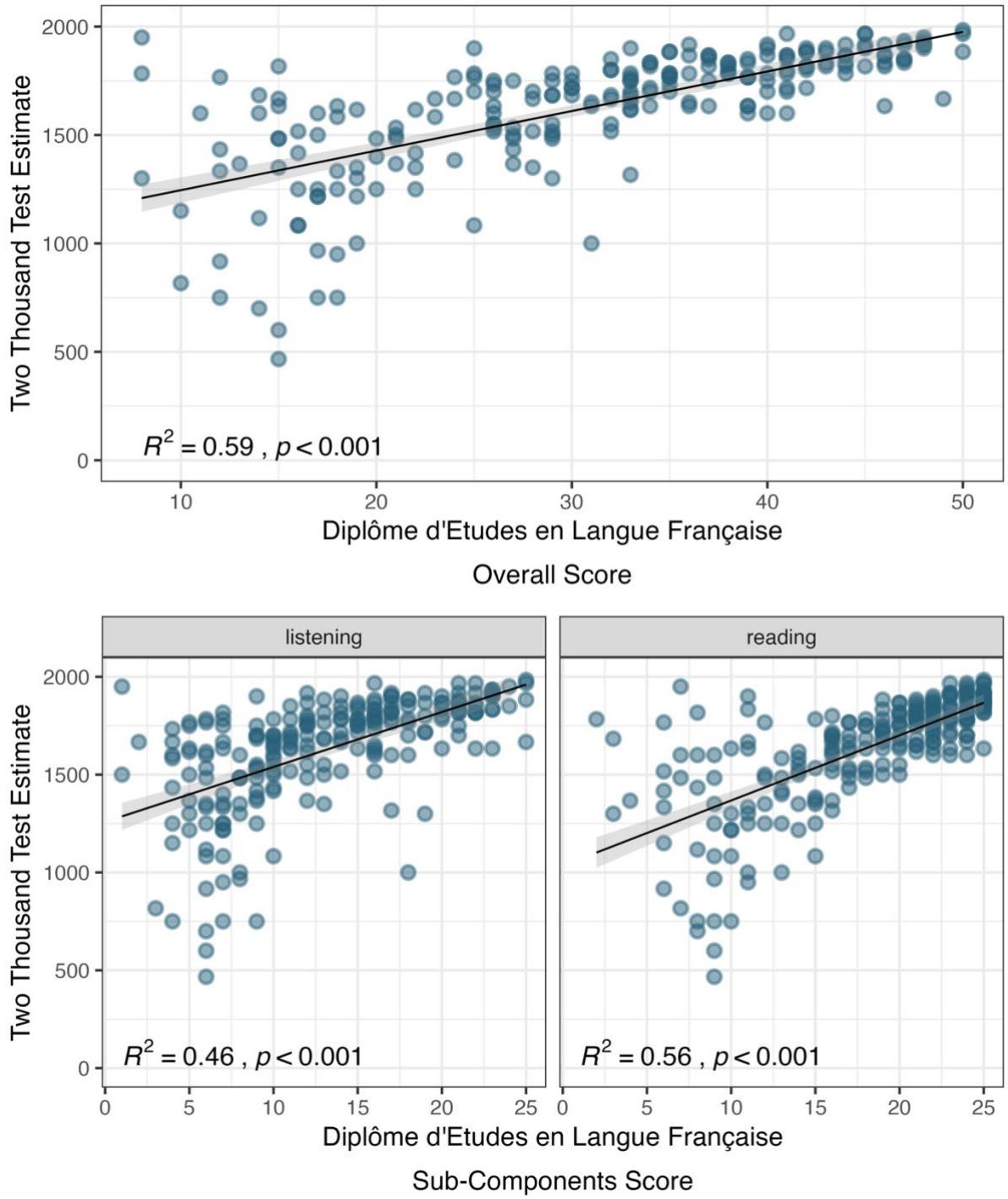

Figure 8. Scatterplots showing associations between CA-TTT size estimates and DELF performance.

Discussion

In response to calls for more rigorous test validation and better specification of each test's purpose, including the type of learners and educational contexts for which the test has been developed (Schmitt et al., 2020), the current study sought to pilot a written receptive test of high-frequency vocabulary for adolescent beginner-to-low intermediate learners of French in instructed contexts in England. In doing so, we provided a snapshot of vocabulary knowledge among mostly high-achieving GCSE French learners and its role in accounting for proficiency. We now discuss our findings in the context of an argument-based validation framework.

Domain Description

When addressing the domain description inference, we observed that learners were more likely to know a word in the CA-TTT (but also X-Lex) if it had appeared in the TLU domain. In this study, the TLU domain was the GCSE curriculum, given learners' very limited exposure to the language outside the classroom. When adjusted to reflect the same proportion of words in the 2,000 most frequent words that were on and off the curriculum list, CA-TTT estimates were significantly lower than the unadjusted estimates (adjusted: M = 1,489, 95% CI [1,449, 1,529]; unadjusted: M = 1,625, 95% CI [1,589, 1,664]. In other words, exposure from *instruction* (classroom, textbooks, homework, etc.) strongly determined vocabulary knowledge. This finding potentially aligns with a recent study (Robles-García et al., 2023) showing that *subjective* exposure—such as teacher judgments about what words students are most likely to know—can moderate vocabulary tests scores as much as—if not more than—the frequency of words in the (arguably more natural and representative) *general* language as a whole.

Our study, although designed to assess knowledge of context-aligned, high-frequency vocabulary, has broader implications for future vocabulary test development. That is, any vocabulary test that randomly samples 20 or even 30 words from each frequency band will

inevitably under- or over-estimate vocabulary knowledge depending on *which* words are selected. Future development of vocabulary knowledge measures could seek a better balance between words that learners could be expected to know (due to their inclusion in the curriculum, for instance) and words that reflect a wider breadth (size) of knowledge (*if* it is of interest to gauge impacts of any out-of-school exposure). However, ascertaining such a balance constitutes a serious challenge: How can we extract meaningful data about 'size' from a relatively small set of words in contexts where exposure to the language is, for many learners, limited to instructed experience? The steps we adopted in the current study may go some way to addressing this challenge, such as adjusting for on- and off-curriculum words. Nevertheless, caution is needed when interpreting vocabulary size estimates in highly instructed, low-exposure contexts.

Generalisation and Scoring Inference

When addressing the generalisation and scoring inference, we found that the CA-TTT measured a unidimensional (i.e., a single underlying) construct, which, we assume, is the construct of form–meaning recognition of vocabulary. Moreover, omega reliability coefficients were high. Perhaps unsurprisingly, given the predictive role of the curriculum and the CA-TTT's high overlap with the curriculum, scores were very high and negatively skewed in both frequency bands. This skew resulted from ceiling effects: Logit estimates for items and persons showed that most items were easy for most participants, with mean person ability above maximum item difficulty. Accordingly, Rasch person separation reliability for the 1,000 band was on the threshold between low and acceptable discrimination, although reliability for the 2,000 band was above the threshold. This suggests that items in the 1,000 band were not as effective at discriminating between different abilities as items in the 2,000 band, at least among our participants.

Item fit was generally satisfactory. Of the 120 items included in the CA-TTT, we identified four items (*éducation*, *femme*, *semaine*, and *puissant*) with poor fit in the *eRm* model estimations. Normally, poorly fitting items would be candidates for substitution during the validation process. However, there are reasons—in addition to their high relevance in the TLU domain—for retaining them. First, fit values obtained for the same items (with the exception of *puissant*) using a different method (*TAM*) were within the optimal range. Second, a closer inspection of these items suggested that poor fit may have resulted from ceiling effects: Three of the four items (*éducation, femme,* and *semaine*) had very low difficulty values, making their outfit mean-square statistics particularly sensitive to mistakes by more-proficient learners who represented a significant proportion of our sample (see below). Administering the CA-TTT with a different sample (e.g., of a lower-proficiency) from the same population could give different results.

Explanation Inference

To assess the explanation inference, we correlated CA-TTT scores with an existing measure of vocabulary size (X-Lex). Although the CA-TTT assesses form–meaning recognition and X-Lex form recognition, we observed a strong and significant positive correlation between CA-TTT and X-Lex scores. This suggests that (a) they are tapping into similar underlying constructs (i.e., form[–meaning] recognition) and (b) the kind of knowledge elicited by one test tends to improve with the kind of knowledge elicited by the other.

However, vocabulary knowledge estimates were different across the tests: CA-TTT estimates (M = 1,624, 95% CI [1586, 1662]) were often two or three times larger than the corresponding X-Lex estimates (M = 711, 95% CI [666, 756]). This indicates systematic differences between the two tests. It could be argued that X-Lex measures a different construct (form recognition) from the CA-TTT (form–meaning recognition) and as such, we should not

expect to see similar scores. Nevertheless, we should expect higher scores on the 'easier' test (X-Lex) than the 'harder' one (CA-TTT). Instead, we see the opposite.

We suggest that these differences can largely be attributed to the number and type of words in the two tests. First, and perhaps most importantly, the CA-TTT contained a far greater proportion of words sampled from the GCSE curriculum list than X-Lex. This, together with the predictive role of the curriculum, is very likely to strongly—or perhaps even entirely—explain differences in the scores obtained by the two tests.

Second, the CA-TTT included 60 items in each frequency band, whereas X-Lex only included 20 items. Stoeckel et al. (2021, p.198) highlight that "the scale of uncertainty" associated with vocabulary size and levels tests (such as X-Lex) is "simply too large for test users to have confidence in such determinations". One way to partially address this "scale of uncertainty" and improve the accuracy of these tests, as suggested by Gyllstad et al. (2021), is to increase the number of target items to *at least* 30 in each frequency band, as we have done for the CA-TTT.

Finally, X-Lex and the CA-TTT sampled from two different frequency lists. X-Lex used an older frequency list (Baudot, 1992) based exclusively on written corpora, whereas the CA-TTT sampled from a more recent list (Lonsdale & Le Bras, 2009) of written *and* spoken materials. Strikingly, frequency values were quite different between the two lists: Of the 40 high-frequency (<2,000) items in X-Lex, only 27 fell within the same frequency band across both the Lonsdale and Le Bras (2009) and Baudot (1992) lists.

Extrapolation Inference

When addressing the extrapolation inference, we found strong associations between (unadjusted) CA-TTT vocabulary size estimates and performance in both high-stakes (GCSE) and standardised (DELF) proficiency tests. Despite a skew towards higher GCSE grades in our

sample (see below for potential reasons), CA-TTT scores correlated strongly with overall and skill-specific (reading, listening, writing, and speaking) GCSE levels. Likewise, CA-TTT scores correlated with DELF listening and reading performance—a test that learners were not familiar with and that contained fewer (high-frequency) words from the GCSE curriculum list (listening: M = 67.94%, 95% CI [57.85%,78.03%]; reading: M = 71.17% [62.77%,79.58%]) and the CA-TTT (listening: 14.17%; reading: 17.50%) than the GCSE exams. As expected, given the written modality of the CA-TTT, the correlations between CA-TTT performance and proficiency were strongest in the reading comparisons, compatible with evidence relating to the strong association between (written receptive) vocabulary knowledge and reading (Jeon & Yamashita, 2022). We also found that correlations between the CA-TTT and proficiency measures overlapped in confidence intervals with those reported in recent meta-analyses (In'nami et al., 2022; Jeon et al., 2022; Jeon & Yamashita, 2022; Kojima et al., 2022; see Appendix S8 for more information). Together, these findings suggest that the reported relations between L2 vocabulary knowledge and proficiency are similar for learners of a language other than English to those found to date for English.

Limitations of the study and future directions

Differences between X-Lex vocabulary size estimates obtained in our study and previous research are noteworthy. The mean estimate in this study was 1,167 (95% CI [1,076, 1,259]), an estimate considerably larger than those previously reported: 852 (Milton, 2006 at the end of Year 11), 775 (Milton, 2015 at the end of Year 11), and 564 (David, 2008 at the beginning of Year 11).

Overall, the percentage achieving level 7 or higher at GCSE in our study (68%) was much higher than the corresponding percentage for the population (31% in 2022 and 26% in 2023; Ofqual, 2023). One reason for the high GCSE performance of our sample (and thus low

discrimination indices) could be self-selection: In our study, teachers told students about the study, but individuals chose to participate. An additional reason could be that although our learners were in the equivalent school year as those in David's and Milton's studies, they were tested *immediately after* their GCSE exams when their knowledge was likely to be strongest.

Interestingly, David (2008) observed a mean estimate of 1,577—only about 500 more words than our study—for students who had received an additional 190 hours of instruction (i.e., in Year 12). David's participants—like many (68%) of ours—had also performed highly at GCSE, with 95% obtaining an A or A* (equivalent now to Level 7 or above). Despite these sampling differences, the fact that our X-Lex scores fell roughly in between the scores observed by David (2008) for Year 11 and Year 12 (564 and 1,577 respectively) suggests that our findings are broadly compatible with those from previous research. Nevertheless, future research should examine the CA-TTT's (preliminary) internal and external validity with participants from a wider (including *lower*) range of knowledge and proficiency to reduce any effects resulting from self-selection bias. Future research could also go a step further in the validation process by ascertaining if the test correlates with entirely different measures, such as grammatical knowledge or phonological awareness, as suggested by Bachman (2004).

Further indicating a skew in our sample was that the percentage of correct answers in the 2,000 band of the CA-TTT (75%) was higher than the performance reported in the TTV validation study by Batista and Horst (69%) for the same frequency band. This is noteworthy, given that Batista and Horst's sample included university students spanning a range of proficiency levels: beginner, low intermediate, high intermediate, and advanced. One explanation for these differences might be that the CA-TTT contained twice the number of items in the 2,000 band than the TTV. Another explanation might be the use of English (rather than French) definitions in the CA-TTT. Size estimates based on bilingual tests have been shown to be larger and more accurate relative to monolingual tests, because they are more

sensitive to partial knowledge especially among beginner-to-low intermediate learners (Elgort, 2013; Nation, 2013). Future research could compare results between the CA-TTT and TTV directly among the same population of learners.

A noteworthy finding from our initial validation was that the test items could be argued to be too 'easy' for our specific sample of learners. As we have argued, this was in large part due to a combination of intentional design features, including the high proportion of words from the curriculum (relative to previous tests used in this context) and the high proportion of high-performing learners. It was critical to test these high-performing learners—given our aims of informing policy and practice about vocabulary knowledge at the end of the GCSE course—but we strongly encourage further validation work with low(er)-proficiency participants at the same stage of education (school year). Such work would build on our assessment of the test's ability to discriminate between individuals, which would be especially important if (a revised) CA-TTT were ever to be used as an achievement test to ascertain students' knowledge as they approach their high-stakes exams.

One intuitive step to address the relative 'ease' of the test, as suggested by an anonymous reviewer, could be to remove words overlapping in difficulty and replace them with low(er)-frequency words, based on an assumption that low(er)-frequency vocabulary is (usually) more difficult than high(er)-frequency vocabulary. However, we argue that, for our highly instructed context, such an approach would most likely be effective in making the test 'more difficult' if these low(er)-frequency words were intentionally *not* from the curriculum list, given the strength of association between the curriculum and vocabulary knowledge observed in our study. We also reiterate that sampling words from low(er)-frequency bands would have run counter to the initial aim of the current study: to test knowledge of, specifically, high-frequency words. To preserve this aim, a more appropriate solution would be to test *a greater number of* high-frequency words on and off the curriculum list in a more balanced

manner or possibly even every word via a bootstrapping methodology, whereby "cases, once sampled, are returned to the population before sampling occurs again" (McLean et al., 2020, p.395). It could be that the words we selected were among the easiest of high-frequency words. Therefore, testing the whole set would allow researchers to determine whether certain high-frequency words are more difficult than others due to factors (beyond frequency alone), such as "semantic neutrality, length, part of speech, polysemy, morphological regularity, cognateness, [and] orthographic transparency" (Hashimoto, 2021, p.182).

Conclusion

The current study extends researchers' and teachers' toolkits by providing information about the internal and external validity of a new, freely available instrument (the CA-TTT) to test context-aligned, high-frequency French vocabulary size for beginner-to-low intermediate proficiency levels in instructed contexts. Preliminary results are promising: The CA-TTT showed high internal and external validity, with scores strongly and positively correlating with another measure of vocabulary size and both standardised and high-stakes proficiency measures. The CA-TTT, once piloted with lower-proficiency learners at the same stage of education and revised as appropriate, could potentially serve as a tool for assessing high-frequency L2 French vocabulary knowledge for students about to take GCSEs, and even as a potential (albeit crude) proxy for proficiency at beginner-to-low intermediate levels at this stage of education.

We do, however, advocate caution when interpreting estimates from vocabulary size tests, including our own, and especially in instructed contexts. In our study, we found that the curriculum played a decisive role in predicting vocabulary knowledge and may have contributed to under-estimations (in the case of X-Lex) or overestimations (in the case of the CA-TTT) of vocabulary size. Thus, without careful consideration of the curriculum context,

such tests could inevitably under- or over-estimate vocabulary knowledge as a function of the relationship between the lexicons of the curriculum and the test. Our study has demonstrated that when designing such size tests and when calculating and interpreting the estimates, it is important to consider the tests' intended purpose(s) and acknowledge an inevitable conflation of vocabulary size tests and achievement tests in highly instructed populations of L2 learners.

Finally, the open accessibility of the tool can, we hope, widen the scope of research producers and consumers (Marsden & Morgan-Short, 2023), adding to the numerous options already available in English. We hope that the CA-TTT inspires the development of equivalent tests for other languages and proficiency levels thus far underrepresented in the literature.

References

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

Aryadoust, V., Ng, L., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Ayres-Bennett, W., Hafner, M., Dufresne, E., & Yerushalmi, E. (2022). *The economic value to the UK of speaking other languages*. RAND. https://doi.org/10.7249/RRA1814-1

Bachman, L. (2004). *Statistical analyses for language assessment book*. Cambridge University Press.

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions*, *21*(3), 1105-1106.

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

Baudot, J. (1992). *Fréquence d'utilisation des mots en français écrit contemporain*. Les Presses de l'Université de Montréal.

Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*, *70*(1), 11-47. https://doi.org/10.1111/lang.12368

Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2022). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, *43*(3), 596-602. https://doi.org/10.1093/applin/amaa061

Chapelle, C. (2012). Validity argument for language assessment: The framework is simple…. *Language Testing*, *29*(1), 19-27. https://doi.org/10.1177/0265532211417211

Coxhead, A., Nation, P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies*, *50*(1), 121-135. https://doi.org/10.1007/s40841-015-0002-3

Dang, T., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*(1), 66-76. https://doi.org/10.1016/j.esp.2013.08.001

David, A. (2008). Vocabulary breadth in French L2 learners. *The Language Learning Journal*, *36*(2), 167-180. https://doi.org/10.1080/09571730802389991

Department for Education. (2022). *GCSE French, German and Spanish subject content*. https://www.gov.uk/government/publications/gcse-french-german-and-spanish-subject-content

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*(2), 253-272. https://doi.org/10.1177/0265532212459028

Finlayson, N., Marsden, E., & Hawkes, R. (2024). A new vocabulary list for beginner-low-intermediate learners of French, German and Spanish aged 11-16. *OSF Preprints*. https://doi.org/10.31219/osf.io/ya9kh.

FLLOC. (n.d.). *French X-Lex 1*. Retrieved 08/03/2022, from http://www.flloc.soton.ac.uk/documents/French_X-Lex_1.pdf

Flora, D. (2020). Your coefficient alpha is probably wrong, but which coefficient omega Is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484-501. https://doi.org/10.1177/2515245920951747

France Éducation International. (n.d.). *DELF junior/scolaire*. Retrieved 21/12/2022, from https://www.france-education-international.fr/en/diplome/delf-junior-scolaire?langue=en

GO FAIR. (n.d.). *FAIR principles*. Retrieved 08/03/2022, from https://www.go-fair.org/fair-principles/

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

Hashimoto, B. (2021). Is frequency enough?: The frequency model in vocabulary size testing. *Language Assessment Quarterly*, *18*(2), 171-187. https://doi.org/10.1080/15434303.2020.1860058

He, X., & Godfroid, A. (2019). Choosing words to teach: A novel method for vocabulary selection and its practical application. *TESOL Quarterly*, *53*(2), 348-371. https://doi.org/10.1002/tesq.483

Hu, H.-C., & Nation, I. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403-430. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/c5873d5c-23b5-41d1-99a5-fde539883ceb/content

Hu, L., & Bentler, P. (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

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

IRIS. (n.d.). *IRIS database*. Retrieved 08/02/2022, from https://iris-database.org/

Jeon, E., In'nami, Y., & Koizumi, R. (2022). Chapter 11. L2 speaking and its external correlates. In E. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency (pp*.339-367). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.13.11jeo

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

Joint Council for Qualifications. (2023). *GCSE Results 2023*. https://www.jcq.org.uk/wp-content/uploads/2023/08/GCSE-FC-Cat-1-Summary-June-2023-v1.1.xlsx

Joint Council for Qualifications. (2003). *GCSE Results 2003*. https://www.jcq.org.uk/wp-content/uploads/2018/11/GCSE-Entry-Level-GNVQ-Results-Summer-2003.pdf

Jorgensen, T., Pornprasertmanit, S., Schoemann, A., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modelling*. https://cran.r-project.org/web/packages/semTools/index.html

Kane, M. (2006). Validation. In *Educational measurement* (4th ed., pp.17-64). American Council of Education and Praeger Series on Higher Education.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000.

Karami, H. (2012). The development and validation of a bilingual version of the vocabulary size test. *RELC Journal*, *43*(1), 53-67. https://doi.org/10.1177/0033688212439359

Kojima, M., In'nami, Y., & Kaneta, T. (2022). Chapter 6. L2 writing and its external correlates. In E. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency (pp*.159-211). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.13.06koj

Kremmel, B. (2021). Selling the (word) family silver? *Studies in Second Language Acquisition*, *43*(5), 962-964. https://doi.org/10.1017/S0272263121000693

Laufer, B., & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15-30. https://files.eric.ed.gov/fulltext/EJ887873.pdf

Linacre, J. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, *16*(2), 878. https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. (2021). R statistics Rasch packages: A survey. *Rasch Measurement Transactions*, *34*(1), 1805-1807. https://www.rasch.org/rmt/rmt341.pdf

Liu, M., Chong, S., Marsden, E., McManus, K., Morgan-Short, K., Al-Hoorie, A., Plonsky, L., Bolibaugh, C., Hiver, P., Winke, P., Huensch, A., & Hui, B. (2022). Open scholarship in applied linguistics: What, why, and how. *Language Teaching*, 1-6. https://doi.org/10.1017/S0261444822000349

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

Mair, P., Hatzinger, R., & Maier, M. (2021). *eRm: Extended Rasch modelling*. https://cran.r-project.org/web/packages/eRm/index.html

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

Marsden, E., & Morgan-Short, K. (2023). (Why) are open research practices the future for the study of language learning? *Language Learning*, *75*(Jubilee). https://doi.org/10.1111/lang.12568

McLean, S., Stewart, J., & Batty, A. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*(3), 389-411. https://doi.org/10.1177/0265532219898380

Meara, P., & Milton, J. (2003). *X_Lex: The Swansea vocabulary levels test*. Express Publishing.

Meara, P., & Miralpeix, I. (2006). *Y_Lex: The Swansea advanced vocabulary levels test.* Lognostics.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp.13-103). Macmillan.

Milton, J. (2006). Language lite? Learning French vocabulary in school. *Journal of French Language Studies*, *16*(2), 187-205. https://doi.org/10.1017/S0959269506002420

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Multilingual Matters.

Milton, J. (2015). French lexis and formal exams in the British foreign language classroom. *Revue Française de Linguistique Appliquée*, *20*(1), 107-119. https://doi.org/10.3917/rfla.201.0107

Nation, P. (1990). *Teaching and learning vocabulary*. Newbury House.

Nation, P. (2013). Testing vocabulary knowledge and use. In P. Nation (Ed.), *Learning vocabulary in another language* (pp.514-568). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656.015

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9-13. http://www.jalt-publications.org/archive/tlt/2007/07_2007TLT.pdf

Nguyen, L., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86-99. https://doi.org/10.1177/0033688210390264

Nicklin, C., & Vitta, J. (2022). Assessing Rasch measurement estimation methods across R packages with yes/no vocabulary test data. *Language Testing*, 39(4), 513-540. https://doi.org/10.1177/02655322211066822

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Ofqual. (2023). *Relationships between grade distributions for students taking different combinations of GCSE subjects*. https://analytics.ofqual.gov.uk/apps/GCSE/9to1/

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

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

Peters, E., Velghe, T., & Van Rompaey, T. (2019). The VocabLab tests: The development of an English and French vocabulary test. *ITL - International Journal of Applied Linguistics*, *170*(1), 53-78. https://doi.org/10.1075/itl.17029.pet

Plonsky, L. (2015). Statistical power, p values, descriptive statistics, and effect sizes: A 'back-to-basics' approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp.23-45). Routledge.

Purpura, J., Brown, J., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, *65*(S1), 37-75. https://doi.org/10.1111/lang.12112

Qualtrics. (n.d.). *Qualtrics*. Retrieved 08/03/2022, from https://www.qualtrics.com

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research* (2.4.1).

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules*. https://cran.r-project.org/web/packages/TAM/index.html

Robles-García, P., Stewart, J., Nicklin, C., Vitta, J. P., McLean, S., & Kramer, B. (2023). 'The wisdom of crowds': When teacher judgments outperform word-frequency as a predictor of students' vocabulary knowledge. *Language Teaching Research*. https://doi.org/10.1177/13621688231176067

Rosseel, Y. (2023). lavaan: Latent variable analysis. In *Journal of Statistical Software* (Vol. 48, Issue 2). https://cran.r-project.org/web/packages/lavaan/index.html

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*(1), 26-43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109-120. https://doi.org/10.1017/S0261444819000326

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55-88. https://doi.org/10.1177/026553220101800103

Stewart, J., Gyllstad, H., Nicklin, C., & McLean, S. (2023). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Language Testing*. https://doi.org/10.1177/02655322231162853

Stoeckel, T., & Bennett, P. (2015). A test of the new General Service List. *Vocabulary Learning and Instruction*, *4*, 1-8. http://vli-journal.org/wp/vli-v04-1-2187-2759/

Stoeckel, T., Bennett, P., & Mclean, S. (2016). Is “I Don’t Know” a viable answer choice on the vocabulary size test? *TESOL Quarterly*, *50*(4), 965-975. https://doi.org/10.1002/tesq.325

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43*(1), 181-203. https://doi.org/10.1017/S027226312000025X

Webb, S. (2021). The lemma dilemma: How should words be operationalized in research and pedagogy? *Studies in Second Language Acquisition*, *43*(5), 941-949. https://doi.org/10.1017/S0272263121000784

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., & Rodgers, M. (2009). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335-366. https://doi.org/10.1111/j.1467-9922.2009.00509.x

Webb, S., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal*, *44*(3), 263-277. https://doi.org/10.1177/0033688213500582

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL - International Journal of Applied Linguistics*, *168*(1), 33-69. https://doi.org/10.1075/itl.168.1.02web

Wright, B., & Linacre, J. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370-371. https://www.rasch.org/rmt/rmt83b.htm

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

Zhang, X. (2013). The I don’t know option in the vocabulary size test. *TESOL Quarterly*, *47*(4), 790-811. https://doi.org/10.1002/tesq.98

[1] We do not report Chronbach's alpha, as per Batista and Horst (2016), as our data violated the tau equivalence assumption (i.e., all items equally loaded onto the same underlying construct).