Received: 21 February 2023 | Accepted: 4 January 2024

DOI: 10.1111/flan.12751

ORIGINAL ARTICLE

ACTFL Language Connects WILEY

The lexical content of high-stakes national exams in French, German, and Spanish in England

Amber Dudley | **Emma Marsden**

The Challenge

Every year, approximately 250,000 16-year-olds in England sit high-stakes exams in French, German, and Spanish. But how many and what kinds of words do these learners need to know to understand the listening and reading exam texts? And how often do these words change year-on-year? This article aims to address these questions by analyzing a corpus of exam papers.

Department of Education, University of York, York, UK

Correspondence
Amber Dudley, Department of Education, University of York, York, UK.
Email: amber.dudley@york.ac.uk

Funding information
Department for Education for England; Research England; Higher Education Innovation Funding; Economic and Social Research Council Impact Acceleration Account; University of York

Abstract

Surprisingly, little is known about the number and frequency level of words that beginner-to-low-intermediate 16-year-old learners of French, German, and Spanish are expected to know when taking high-stakes national exams in England. This study presents exploratory analyses of the lexical content of the listening and reading tests of these exams, a corpus totaling 116,647 running words. Specifically, it seeks to understand the number and frequency level of words that (a) this demographic seems to be expected to know and (b) could be needed for awarding organizations to create exams year-on-year. Key findings include that the proportion of low(er)-frequency words in the corpus of exam papers seemed large, given the stage of the learners and the purpose of the assessments.

Critically, these low(er)-frequency words changed at a high rate between papers, likely incurring a heavy reliance on the lexical inferencing abilities of these relatively inexperienced language learners.

KEYWORDS
French, German, high-stakes exams, Spanish, word frequency

1 | INTRODUCTION: EDUCATIONAL CONTEXT IN ENGLAND

Vocabulary knowledge is reported to strongly predict second language (L2) listening and reading (Zhang & Zhang, 2022). This is not surprising. Research has shown that learners need to know at least 95% of the words in any given written or spoken text to be able to fully understand it (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). It has therefore been recommended that language educators prioritize the words that are most useful to learners (Laufer & Nation, 2012; Nation, 2013), especially during the early stages of learning when both the language system (i.e., knowledge and skills) and motivation are fragile. High-frequency words are considered among the most useful given that a relatively small set of (approximately 2,000) words cover a significant proportion (82%– 89%) of written and spoken language in English (Dang & Webb, 2014; Webb & Nation, 2017).

Despite the importance of vocabulary knowledge, very little is known about the lexical content of the General Certificate in Secondary Education (GCSE) in foreign languages. GCSEs are high-stakes qualifications taken by approximately 600,000 16-year-olds in England every year,[1] with performance playing a critical role in individuals' future academic trajectory and job prospects. Most students take GCSEs in the core (compulsory) subjects—Maths, English, and Science—as well as four or five optional subjects. For 45.95% of the 2023 GCSE cohort, one of these optional subjects was a language (Department for Education, 2023).

The GCSE exams must assess the "subject content" set out by the government's Department for Education. The subject content outlines the knowledge, understanding, and skills that must be tested. For foreign languages, these exams include skill-based proficiency tests in listening, reading, speaking, and writing and are taken after approximately 400-to-450 hours of classroom instruction. The exams are developed by awarding organizations (e.g., AQA [Assessment and Qualifications Alliance] and Pearson Edexcel) whose activities are regulated by the Office of Qualifications and Examinations Regulation (Ofqual) to make sure that qualifications are awarded consistently and fairly. Awarding organizations are commercial entities: Schools choose which organization they enter their students for,[2] and exam entry fees are paid by schools (i.e., via the tax-payer for government-funded schools attended by approximately 93% of English schoolchildren).

Before the GCSE exams, schools select an "entry tier" for each student based on prior attainment. Lower-achieving students are entered for foundation tier (for levels [grades] 1–5), and higher-achieving students for higher tier (for levels 4–9). In 2023, 69.92% of students entered for a GCSE language achieved a standard pass (i.e., Level 4 or above) and 55.12% a good pass (i.e., Level 5 or above; Department for Education, 2023). Level 4 is approximately

equivalent to A1 (with some exams in some languages A1/A2) on the Common European Framework of Reference for Languages scale (Curcin & Black, 2019) and to Intermediate Low/Mid on the ACTFL Proficiency scale (ACTFL, n.d.).

In recent years, there has been a downward trend in the numbers studying languages. Between 1992 and 2004, all 14-to-16-year-olds were required to study at least one foreign language (The Education National Curriculum Modern Foreign Languages Order, 1991). Following the removal of this requirement, announced in 2002, uptake (the number of students taking a language) in England declined by almost 50% (Churchward, 2019), and in 2023, only 45.95% of 16-year-olds studied a language (Department for Education, 2023). Family background is an important predictor of uptake: In 2023, 34.31% of disadvantaged children took a GCSE language, compared with 49.19% of their non-disadvantaged peers (Department for Education, 2023).

This overall decline has often been linked to the perceived and evidenced difficulty of the language exams. Several studies (Coffey, 2016; Parrish & Lanvers, 2019; Taylor & Marsden, 2014) have associated low motivation and uptake at GCSE (and beyond) with a general perception that languages are more challenging than other subjects. Moreover, an Ofqual report (He & Black, 2019) found that the grading system (i.e., the relationship between raw scores and grades assigned) was more severe for GCSE French and German—but not Spanish—than for other subjects. In other words, top grades were generally harder to achieve in languages relative to other subjects. In response, Ofqual introduced adjustments to grading standards for French and German.

Although such adjustments serve as one step in overcoming a disparity in top grades between foreign languages and other subjects, they cannot tell us about the potential *causes* of the perceived difficulties of foreign languages. Indeed, one motivation for the current exploratory study was to consider whether the nature of the lexicon within the assessments could be contributing, to some extent, to languages being considered one of the most difficult subjects. A second motivation was a government review of the current GCSE foreign language subject content (Department for Education, 2021a), set up in response to the challenges outlined above. A proposal that emerged from that initial review (Department for Education, 2021b, p. 4) was to require a more consistent and standardized treatment of vocabulary, by defining the number and frequency levels of words to be included in new compulsory wordlists from which awarding organizations must sample from when creating exams. At the time of finalizing this article, the Department for Education review had concluded, and informed by some of the findings from the current study, awarding organizations are now preparing new exams for 2026 to assess a revised subject content (Department for Education, 2022).

The present study analyzed the lexical content of the available exams that assessed the *current* GCSE subject content to broaden understanding among researchers, policy-makers, assessment regulators, and test-developers and to inform the Department for Education's review described above. The findings informed early thinking about the size and nature of any potential future wordlist in line with calls from researchers to examine the relevance of wordlists for curricula and assessments (Dang et al., 2020). For our analyses, we calculated the basic characteristics—that is, the volume and frequency bands—of the vocabulary used in the current GCSE listening and reading exams in French, German, and Spanish. Analyzing the listening and reading exams thus provided a proxy for the maximum number of words a test-taker would need for the whole GCSE, as receptive knowledge is generally larger than productive knowledge (Webb, 2008).

2 | BACKGROUND LITERATURE

We first outline the *current* context: The stipulations regarding the lexical content of GCSE exams between 2015 and 2025 and the available evidence about the number of words known at these early stages of learning. We then review the research into three key issues that motivated our aims, analyses, and discussion of the lexical content of these exams: (a) the relations between vocabulary knowledge and comprehension across modalities; (b) the importance of word frequency for comprehension; and (c) the implications of word frequency for lexical inferencing.

2.1 | The current context of foreign language education in England

2.1.1 | Policy specifications for the current GCSE in foreign languages

To date, it has been difficult to describe the lexical content of GCSE exams given the absence of (a) specialized lexical profiling software (at least for German and Spanish[3]) and (b) clear guidance about vocabulary. Regarding the latter, the Department for Education's (2015) current subject content—operational between 2015 and 2025—describes the *grammar* requirements in its appendices, but there is little reference to the type or amount of *vocabulary* that the awarding organizations can or should assess. Nevertheless, awarding organizations choose to provide *optional* wordlists. These are structured broadly similarly across organizations, with one section dedicated to general or high-frequency language and another to so-called topic-specific language. Critically, these lists are non-exhaustive and intended only as a guide for teachers and textbook publishers. Although the current subject content does not stipulate whether or how often words from these lists must be included (i.e., sampled) in the exams, the regulatory body (Ofqual, 2021b) asserts that assessments must cover the entire subject content over three-to-five years. Critically, Ofqual (2016) has also required awarding organizations to use words in the exams that were ***not*** on their lists.[4]

Mention of word frequency is very limited both in the subject content and the awarding organizations' exam specifications. For example, AQA (2016, p. 13) makes one explicit reference ("students are expected to be able to [...] understand general and specific details within texts using high-frequency familiar language across a range of contexts") and one implicit reference (i.e., a "general vocabulary" section in the wordlists). Pearson Edexcel (2018, p. 24) provides a list of "high-frequency words" and suggests that examples of creative language use involve "manipulating language, including familiar, high-frequency, and simple language." However, these documents do not specify what "high-frequency" means or how it should be measured. Furthermore, the Department for Education (2015) does not constrain the amount or proportion of high- or low(er)-frequency words that the organizations can use in their exams.

Some tentative evidence suggests that the vocabulary used in the current exams may be contributing—at least in part—to the difficulty of foreign languages relative to other subjects. Stratton and Zanini (2018, p. 5) found that the 2015 reforms to the GCSEs (first examined in 2018) led to an unexpected and undesirable increase in difficulty (as measured by subject experts) "due to an increase in the demand of the vocabulary used in the reading and listening texts" between 2017 and 2018. They reported that lexical variety (i.e., the proportion of unique words) contributed to this increase in difficulty in the reading (but not listening) exams in French and Spanish (but not German). One limitation of that study, however, was its very small

corpus of exams. It only compared two sets of exams from the three leading awarding organizations (AQA, Edexcel, and WJEC): One before and one after the 2015 reforms.

As such, it is not well understood how many different words students are expected to know when taking a GCSE, that is, how many different words might appear in a series of exams over the years. Nor is it known what proportion of words are high-frequency. A better understanding of the lexical content of these exams would help to inform future development of exams and curricula content. In sum, the current study fills an important gap, given the lack of top-down guidance and robust evidence surrounding the lexical content of these high-stakes examinations.

2.1.2 | Vocabulary knowledge of school children in England

Our understanding of how many and what kinds of words students know when they sit their GCSEs is also limited, with just three known studies to date, each with quite small sample sizes and each using the French version of *X-Lex* (Meara & Milton, 2003), a Yes/No (self-report) vocabulary test of form recognition as the sole measure. Milton (2006) estimated that when learners took their GCSE exams, their receptive knowledge consisted of a mean of 852 words ($n = 49$, standard deviation [SD] = 440, range: 0–1,800). Such a large standard deviation and range suggest a skewed distribution below the mean. David (2008) and Milton (2015) observed lower sizes, reporting a mean estimated vocabulary size of 564 ($n = 26$, SD = 352, range: 0–1,650) and 775 ($n = 18$, SD = 341, range: 350–1,250) words, respectively. Moreover, these studies found that learners typically knew more high-frequency words than low(er)-frequency ones, although the proportions of known words were small across all frequency bands.

The instrument used (i.e., X-Lex) in these studies, however, selected test items without consideration of the curriculum. This could be problematic given that the curriculum can heavily determine vocabulary knowledge. Dudley et al. (2024, under review), for instance, estimated 222 GCSE learners' vocabulary size in French using two tests: X-Lex, which has little overlap between test items and curriculum content, and a Context-Aligned Two Thousand Test (CA-TTT), which has a high overlap between test items and the curriculum content. They found that CA-TTT estimates ($M = 1{,}624$, 95% confidence interval [CI]: [1,586–1,662]) were often two or three times larger than the corresponding X-Lex estimates ($M = 711$, 95% CI: [666–756]).

A limitation of all these studies is their focus on receptive vocabulary knowledge using a written form recognition test. They therefore cannot inform us about knowledge in the oral modality, since differences in knowledge can vary as a function of modality given the static and controlled nature of word knowledge in the written modality and the ephemeral nature of the spoken modality (Milton, 2009; Read, 2007; van Zeeland, 2013).

Although we have some broad indications as to how many and what words GCSE students may recognize, no study has systematically analyzed the lexical content of the exams. As such, we know relatively little about how many words learners are *expected* to know, or whether this number differs across languages, years, tiers of entry, or modality. Numerous stakeholders within the language education community, including teachers, teacher educators, test developers, curriculum designers, textbook publishers, and researchers, have therefore had to operate with a limited understanding of the expectations about vocabulary knowledge. One consequence of this for teaching and assessment is that teaching materials may not align with the language that learners actually face in their high-stakes exams. One consequence for

research, as Dudley et al. (2024, under review) argue, is that researchers must consider the curriculum when selecting which words to test among instructed learners, given the curriculum's role in predicting vocabulary knowledge.

2.2 | Relations between vocabulary knowledge and comprehension

Understanding the volume and nature of the lexicon used in listening and reading assessments is crucial given the widely observed positive relationship between vocabulary knowledge and comprehension (Zhang & Zhang, 2022). It has been estimated that learners need to know at least 95%—but optimally 98% for unassisted comprehension of the full text—of the words when reading to engage in higher-level comprehension, although knowing between 55% and 60% may support minimal comprehension (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). Indeed, Kremmel et al. (2023) recently found "a mostly linear relationship between vocabulary coverage and reading comprehension" (p. 31), as per Hu and Nation (2000), although in their nonacademic context, they did not replicate the 98% coverage threshold. Kremmel et al. suggested that the impact of unknown word density on comprehension may vary as a function of "the genre and response format used to measure comprehension" (p. 31).

The level of coverage and, by extension, the number of words needed for comprehension may also vary as a function of modality (Adolphs, 2003; Nation, 2006; Stæhr, 2009; van Zeeland & Schmitt, 2013). For instance, van Zeeland and Schmitt (2013) reported that adequate listening comprehension can be achieved with only 90% coverage (rather than 95% or 98% as found for reading). They do, however, warn that considerable variation can be expected, particularly among beginner and intermediate learners, and so propose that 95% coverage is necessary for relatively good comprehension to avoid this variation.

To date, most lexical coverage research has focused on L2 learners of English. As such, it is difficult to know with certainty the extent to which these findings similarly pertain in languages other than English, such as French, German, and Spanish. It has been suggested that fewer words may be needed in French (than in English) for successful reading comprehension (Milton & Alexiou, 2009). However, little research, as far as we are aware, has empirically explored this hypothesis beyond simply correlating vocabulary size estimates with CEFR levels.

Taken together, these findings, albeit not comprehensive or conclusive, do suggest that analyzing the vocabulary used in the listening and reading exams will give a reasonable proxy for the number of words test-takers can be *expected* to know for unassisted comprehension of the full text in preparation for *any* comprehension question that may be asked.

2.3 | The importance of knowing high-frequency words for comprehension

A separate but closely related factor that affects comprehension is the frequency level of the words in a text (i.e., whether they are among the most frequent words in the language as a whole). By definition, high-frequency words are more likely to occur than low(er)-frequency words. This means that knowing high-frequency words can help—but, of course, does not guarantee—comprehension.

Views differ as to whether the first 2,000 (Nation, 1990, 2013; Read, 2000; Thornbury, 2002) or 3,000 (Schmitt & Schmitt, 2014) most frequent words should be considered "high-frequency." For the purposes of the current analyses, we define high-frequency as the first 2,000 most frequent words given (a) our focus on the initial stages of learning; (b) evidence that the vocabulary size of students at this level may average some way below 1,000 words (David, 2008; Milton, 2006, 2015); (c) the limited instruction time available; and (d) observations that the first 2,000 most frequent words account for 82%–89% of the words in a variety of texts (Dang & Webb, 2014; Webb & Nation, 2017). For any word beyond that cut-off, for the sake of simplicity, we refer to it as "low(er)-frequency," while acknowledging that this bracket will contain some high-, mid-, and low-frequency words, depending on the criteria used.

Estimates (Dang & Webb, 2014; Webb & Nation, 2017) suggest that the first 2,000 most frequent words cover at least 82% of (formal) written language (e.g., newspapers, novels) and 89% of spoken language (e.g., conversations, television, films, and lectures) in English. One might therefore reasonably predict a higher proportion of high-frequency words in the listening than reading exams. Although little data exist for languages other than English, a comparable level of coverage could be expected given Zipf's (1935) law. See Cobb and Horst (2004) for evidence that coverage may even be greater in French.

In response to such findings, certain jurisdictions have combined corpus-based word frequency data with stakeholders' judgments about word usefulness to develop wordlists. For instance, Hong Kong's Education Bureau (2021) collaborated with the Chinese University of Hong Kong and primary and secondary school teachers to create wordlists for their English language curriculum based on frequency data from the General Service List (West, 1953), the British National Corpus (Nation, 2012), and the Academic Word List (Coxhead, 2000). Similarly, the State of Israel's Ministry of Education worked with researchers, curriculum designers, teachers, textbook writers, and assessment specialists to design a lexical syllabus for the new English curriculum informed by research findings about vocabulary in textbooks and "gaps between learners' lexical knowledge and the amount of [high-frequency] lexis necessary for performing language tasks" (Laufer, 2023, p. 151). Frequency-informed lists have also been used in the testing and teaching of languages other than English, including in the development of Dutch proficiency exams in the Netherlands (College voor Toetsen en Examens, n.d.) and in the French language curriculum in primary schools in France (éduscol, 2020).

In sum, understanding the frequency levels of the words used in the exams can provide insight into (a) current expectations surrounding the lexicons of beginner-to-low-intermediate 16-year-old learners and (b) the potential difficulty of the current exams, while, critically, *not* simply equating frequency with difficulty—a point we return to in the Conclusion. In turn, this improved understanding has the potential to feed into future decisions about the design of assessments, curricula, and teaching materials.

2.4 | Word frequency and lexical inferencing

Our interest in examining the proportion of high- versus low(-er) frequency words used in exams intended for beginner-to-low-intermediate learners is also motivated by the possible effects of word frequency on lexical inferencing (henceforth, inferencing) and its implications for test-takers. Inferencing refers to the process of "making informed guesses as to the meaning of an utterance in light of all available linguistic cues in combination with the learner's general

knowledge of the world, [their] awareness of context, and [their] relevant linguistic knowledge" (Haastrup, 1991, p. 40).

It is generally expected that learners know more high-frequency than low(er)-frequency words. For instance, research has consistently shown that word frequency and other word-related variables (such as cognateness) moderate word learning and, ultimately, overall L2 abilities (e.g., De Wilde et al. (2020). Problems can therefore arise when texts contain a high proportion of low(er)-frequency words with little-to-no orthographic, phonological, or semantic overlap with the first language (L1). To cope with unknown language, learners must draw on compensatory strategies, including inferencing. However, when reading, L2 learners have been found to infer the meaning of only 21% to 59% of unknown words (see, e.g., Laufer, 2020; Wesche & Paribakht, 2009). Contextual factors, including how often the unknown word occurs in the text, the relevance of the unknown word for comprehension, and the proportion of unknown words in the text (Sternberg, 1987), as well as learner-related factors, such as L2 vocabulary size (e.g., Hatami & Tavakoli, 2013), can also determine lexical inferencing success in reading.

Stronger inferencing skills may also compensate for smaller vocabulary sizes. Laufer (2020), for instance, tested 60 adolescent L2 intermediate learners of English and found that at 95% and 98% coverage (i.e., when unknown words represented 5% and 2% of the text, respectively), learners demonstrated high inferencing success rates. In contrast, at 90% coverage (i.e., when 10% of the words were unknown), learners could only infer half of the unknown words. Despite these differences in coverage and inferencing rates, learners demonstrated similar reading scores. One possible explanation, as argued by Laufer, is that successful inferencing increased coverage to 95%—perhaps the minimum level needed for comprehension on that test—from the original 90%. This suggests that learners with stronger inferencing skills but smaller vocabulary sizes may be able to understand as much as learners with weaker inferencing skills but larger vocabulary sizes, at least when reading.

Few studies, however, have examined inferencing in listening. The available research (van Zeeland, 2014) suggests that inferencing success is considerably lower in listening than reading for several reasons. First, learners may identify fewer unknown words in listening than in reading and therefore have fewer opportunities to use inferencing skills. Second, they may have difficulties understanding the necessary clues to the meaning of unknown words because the sound stream is fleeting, whereas our eyes can revisit parts of written input. Third, cognates are less likely to have a facilitatory effect in listening than reading given that phonological cognateness is often less transparent and prevalent than orthographic cognateness (Lubliner & Hiebert, 2011).

As such, the more low(er)-frequency, non-cognate, and, by extension, unknown vocabulary there is in a text, the greater the chances are that test-takers will find reading and, perhaps especially, listening difficult.

3 | THE CURRENT STUDY

To improve our understanding of the volume and frequency level of the lexical content of high-stakes exams in the main languages (French, German, and Spanish) taught in England, we examined four sets of GCSE listening and reading exams from the two leading awarding organizations (AQA and Pearson Edexcel). Two research questions (RQs) were addressed:

RQ1: *What is the number and frequency level of words used:*

a. in an *average* exam? And does this vary as a function of tier, modality, and year?
b. consistently in *every* exam? And does this vary as a function of tier?
c. *only once* across four sets of exams? And does this vary as a function of tier?

RQ2: *How many high- and low(er)-frequency words are used in the corpus of exams? And to what extent do these words change year-on-year as a function of modality and tier?*

4 | METHODS

4.1 | Description of the corpus of exams

The corpus analyzed consisted of 116,647 (function and content) words from a total of 96 exams from four sets (years) of exams (2018, 2019, 2020, and sample), two awarding organizations (AQA and Edexcel), three languages (French, German, and Spanish), two tiers (foundation and higher), and two modalities (listening and reading). The exams contained texts of between 10 and 160 words and (multiple-choice and short open-response) questions in English and in the target language. Examples of exam texts and comprehension questions are available on the awarding organizations' websites (AQA, 2023; Pearson Edexcel, 2023)

This corpus represented all the exams published by the two awarding organizations for the Department for Education's (2015) subject content during our analysis period (October-to-December 2021), as exams were not produced in any subject in 2021 due to the pandemic. (Exams were developed in 2020 but not taken by students.) Each organization's sample exam was included within this corpus as parity (which our analyses checked) seemed likely, given that the sample exam often informs a school's choice of awarding organization. It is therefore in each organization's commercial interests that their sample exams align with those taken by students.

4.2 | Data preparation

Before profiling each exam, we removed any rubrics or instructions in English. Proper nouns were included in the analyses, since they are often ascribed frequency values in corpora (Kilgarriff et al., 2014); they often have a dictionary entry (e.g., countries and cities); and their meaning is not always transparent to learners. Compound nouns in German (when two or more nouns are joined together to create a new word) were split using CharSplit (Tuggener, 2016), an *n-gram*-based compound splitter for German, but only when they were not entries in the FreeLing 3.0 dictionary (Padró & Stanilovsky, 2012). Compounds were not split in French or Spanish because: (a) compounds are not easily distinguishable from multi-word units in French and Spanish and (b) most compounds (and multi-word units) in French and Spanish are highly lexicalized and thus have dictionary entries (van Goethem & Amiot, 2019).

4.3 | The lexical profiler

The listening and reading exams were profiled individually and then in sets grouped according to year, using the MultilingProfiler (http://multilingprofiler.net/; Finlayson et al., 2022, 2023).

To profile each text, we selected a frequency list, a language (French, German, or Spanish), and the level (top 5,000) from the drop-down menus; pasted the exam texts into the profile window; and then pressed "Download Stats (.csv)" to obtain lists (in.csv format) of the flemmas and their members used in the profiled text.

Flemmas consist of a headword and their inflections but, unlike lemmas, do not take the part of speech into consideration (Bauer & Nation, 1993; Webb, 2021). For instance, *sourire* (smile) and *sourires* (smiles) are members of one noun lemma, and *sourire* (to smile) and inflected forms (e.g., *souris* [smiles], *souri* [smiled]) are members of a verb lemma. Both of these lemmas represent one flemma. Although coverage has typically been measured in terms of word families (including inflectional and derivational morphological forms), we chose the flemma as the most appropriate lexical unit for two reasons. First, GCSE students (i.e., beginner-to-low-intermediate learners) are unlikely to possess the relevant knowledge to comprehend all the derivational forms of known headwords (Brown et al., 2022; Cobb & Laufer, 2021). Thus, using word families would have risked overestimating the coverage that high-frequency words provided of the exams. Second, the MultilingProfiler does not currently support part-of-speech tagging and is therefore flemma-based. It is an empirical question—beyond the scope of the current study—whether the (f)lemma distinction significantly impacts coverage calculations (Kremmel, 2021; Webb, 2021).

The .csv files contained information about the flemmas and their inflected forms, including their number of occurrences within the exams and frequency band (0–1,000, 1,001–2,000, 2,001–3,000, 3,001–4,000, 4,001–5,000, and >5,000). These frequency bands were based on corpora-informed frequency lists of the 5,000 most frequently occurring flemmas in the respective languages (Davies & Davies, 2017; Lonsdale & Le Bras, 2009; Tschirner & Möhring, 2019).

These large, general corpora include written and spoken language from a wide range of genres and domains of use. Research, including our own, has found very substantial overlap between different large, general corpora in terms of the information that they render about the most highly frequent words (see, e.g., Brezina & Gablasova, 2015). Using corpora specific to particular domains of use was inappropriate for several reasons. First, there are very few corpora in *all three* languages that use sufficiently similar methods of compilation to allow for meaningful comparisons *between* the languages. As such, the choice was limited when it came to selecting which corpora could provide frequency information for exploring the exams' lexicon. Second, creating corpora for *all three* languages would have been incredibly time-consuming and complex, given the lack of evidence and agreement about which domains of language use these learners *might* need to operate in one day (if ever).

4.4 | Analysis

To address RQ1, we examined the number and frequency level of words used in an average exam (RQ1a) and words from flemmas used in every set of exams (RQ1b) and only once across four sets of exams (RQ1c). See the R scripts on our OSF repository for how the profiling output was manipulated to create these datasets. For these analyses, we calculated descriptive statistics (means and SDs) and ordinal regression models for each language and awarding organization. To compute these models, we used the *clm* function from the *ordinal* package (Christensen, 2019), with the six-level ordinal variable—frequency band (0–1,000, 1,001–2,000, 2,001–3,000, 3,001–4,000, 4,001–5,000, >5,000)—as the dependent variable. Within each model, each unique flemma was weighted by how many times its members occurred in the relevant set of exams to model the coverage of the text provided by the flemma *and* its members.

Tier, modality, year, and their three-way interaction were included as predictors for the RQ1a (words used in an average exam) models. For any model that included a two- or three-way interaction, we only report the highest order interaction as this supersedes the significance of any related main effects and/or lower order interactions. For the RQ1b (word from flemmas used in all four sets of exams) and RQ1c (word from flemmas used in only one set of exams) models, we treated the listening and reading exams as one combined exam and thus only included tier as a predictor.[5]

All predictors were ANOVA-coded, using the *contr_code_anova()* function from the *faux* package (DeBruine et al., 2021), with the intercept set as the grand mean. For tier, we set foundation as the reference level and for modality, reading as the reference level. For year, the sample exam was set as the reference level to investigate whether the sample exam was representative of an actual exam taken by students.

We acknowledge that the level of coverage required for successful comprehension may vary as a function of language. For instance, Cobb and Horst (2004) reported that the 2,000 most frequent words may provide greater coverage of written texts in French than in English. However, given that different corpora and frequency lists were used for each language, and each language has different morphosyntactic features that affect coverage, it made theoretical sense to model each language separately. Moreover, adding language would have produced highly complex four-way interactions that would have been difficult to interpret.

Model summaries were calculated using the *tab_model()* function from the *sjPlot* package (Lüdecke, 2021) and included odd ratios (ORs), 95% confidence intervals (CIs), *p* values for each predictor, and Nagelkerke's pseudo R^2 to assess the fit of the model. Significance was evaluated at an alpha level of .05. Where relevant, *emmeans* (Lenth, 2021) was used to probe any significant interactions. Provided in parentheses are ORs, including 95% CIs and *p* values, with full model summaries reported in Appendix S2. Caution must be exercised, however, when interpreting any significant differences between modalities, tiers, and years, given the low Nagelkerke's R^2 of each model. It is likely that factors (not investigated in this study, including randomness) also contributed to the proportions of high-frequency words used.

To address RQ2, we first calculated the number of high- and low(er)-frequency flemmas in the corpus of exams. We then analyzed the extent to which these flemmas changed year-on-year to explore how predictable (i.e., likely to reoccur) low(er)-frequency flemmas were relative to high-frequency flemmas. For these analyses, "indices of variability" were computed by dividing *the number of unique flemmas used in all four sets of exams* (i.e., 2018, 2019, 2020, and sample exams) by *the sum of the number of flemmas used in each of the four sets of exams*. That is, the numerator counted each flemma once, whereas the denominator included duplicates if a flemma appeared in more than one set of exams. These analyses also provided critical information about the number of flemmas that were awarding organizations needed to create four sets of exams.

Due to space constraints, the Edexcel analyses are only presented in the manuscript, where findings are patterned differently from the AQA analyses. Data sets, analyses, and appendices are provided on our OSF site (https://osf.io/kmn39/).

5 | RESULTS

5.1 | Number and frequency level of words used in an average exam, and whether this varies as a function of tier, modality, and year (RQ1a)

Table 1 shows that an average AQA exam (where listening and reading were combined) had a mean of 614 unique flemmas at foundation and 788 at higher, when averaged across the three

TABLE 1 Mean (SD) cumulative coverage statistics for words from flemmas used in an average AQA exam (listening and reading combined).

	French		German		Spanish		Mean	
	Foundation	**Higher**	**Foundation**	**Higher**	**Foundation**	**Higher**	**Foundation**	**Higher**
0–1,000	76% (2%)	78% (0%)	80% (0%)	80% (1%)	80% (2%)	82% (1%)	78% (2%)	80% (2%)
1,001–2,000	83% (1%)	85% (0%)	87% (0%)	86% (1%)	87% (1%)	89% (1%)	86% (2%)	87% (2%)
2,001–3,000	88% (1%)	89% (0%)	90% (0%)	90% (0%)	90% (0%)	92% (1%)	89% (1%)	90% (1%)
3,001–4,000	90% (1%)	91% (0%)	92% (0%)	92% (0%)	92% (1%)	94% (0%)	91% (1%)	92% (1%)
4,001–5,000	94% (0%)	95% (0%)	93% (0%)	93% (0%)	93% (1%)	95% (0%)	93% (0%)	94% (1%)
>5,000	100% (0%)	100% (0%)	100% (0%)	100% (0%)	100% (0%)	100% (0%)	100% (0%)	100% (0%)
Total no. of words	2,267 (320)	3,208 (418)	1,957 (195)	2,732 (254)	1,876 (166)	2,745 (153)	2,033 (277)	2,895 (353)
Total no. of flemmas	645 (34)	812 (54)	577 (44)	755 (44)	620 (28)	798 (32)	614 (44)	788 (47)

languages and four sets of exams. On average, across the three languages, 86% of (running) words at foundation tier and 87% at higher were high-frequency.

The ordinal models revealed very mixed findings and no clear patterns in terms of variation in the proportion of high-frequency words as a function of modality, tier, or year. The proportion of high-frequency words varied as a function of modality for the AQA German exams and in interaction with tier and/or year for the AQA French and Spanish and the Edexcel French exams. In 15 cases, the listening exams had a greater proportion of high-frequency words than the reading exams, as might be expected. These cases included the AQA German exams at foundation and higher tier (OR: 1.24, 95% CI: [1.15–1.34], $p < .001$), the AQA Spanish and Edexcel French exams at foundation and higher tier in 2020 (OR: AQA Spanish: 1.25, 95% CI: [1.07–1.45], $p = .005$; Edexcel French: 2.13, 95% CI: [1.95–2.36], $p < .001$), and the AQA French exams at higher tier in 2018 (OR: 1.55, 95% CI: [1.31–1.83], $p < .001$) and foundation and higher tier in 2019 (OR: foundation: 1.30, 95% CI: [1.06–1.59], $p = .012$; higher: 1.26, 95% CI: [1.07–1.48], $p = .007$). In contrast, in five cases, the reading exams had a greater proportion of high-frequency words than the listening exams. These cases included the AQA Spanish exams in the sample at foundation and higher tier (OR: 1.21, 95% CI: [1.03–1.41], $p = .019$); the Edexcel German exams in 2018 at foundation and higher tier (OR: 1.31, 95% CI: [1.15–1.50], $p < .001$), and the AQA French exams in the sample at higher tier (OR: 1.61, 95% CI: [1.38–1.89], $p < .001$). In the other 28 cases, the reading and listening exams had similar proportions of high-frequency words.

Tier moderated the proportion of high-frequency words only in the AQA Spanish and Edexcel French exams, in that, perhaps counter-intuitively, the higher tier exams had a greater proportion of high-frequency words than those at foundation tier (OR: AQA Spanish: 1.18, 95% CI: [1.09–1.28], $p < .001$; Edexcel French: 1.16, 95% CI: [1.08–1.23], $p < .001$).

Year of the exam did not systematically affect proportion of high-frequency words, with just two inconsistencies found between the sample and actual exams (in addition to the modality-dependent ones reported above): the sample AQA German exams had a greater proportion of high-frequency words than their 2020 counterparts (OR: 1.13, 95% CI: [1.02–1.27], $p = .023$). In contrast, the 2019 Edexcel Spanish exams had a greater proportion of high-frequency words than their sample counterparts (OR: 1.24, 95% CI: [1.11–1.39], $p < .001$).

5.2 | Number and frequency level of words used consistently in every exam (RQ1b)

Across the three languages, a mean of 175 flemmas consistently appeared in every AQA exam at foundation tier and 222 at higher (Table 2). These flemmas represented a mean of 69% of all words used in each exam at foundation and 71% at higher, when averaged across the three languages. Almost all words from flemmas used in every exam (97% at foundation and 98% at higher, averaged across the three languages) were high-frequency.

The ordinal models revealed that the proportion of high-frequency words varied as a function of tier, but only for the AQA French and Spanish (not German) exams. Specifically, and perhaps surprisingly, the higher exams had a greater proportion of high-frequency words than the foundation exams (OR: French: 1.25, 95% CI: [1.11–1.41], $p < .001$; Spanish: 1.44, 95% CI: [1.16–1.79], $p = .001$). An identical trend emerged for the Edexcel exams.

TABLE 2 Mean (SD where relevant) cumulative coverage statistics for words from flemmas used in every set of AQA exams (listening and reading exam papers were combined).

	French		German		Spanish		Mean (SD)	
	Foundation	**Higher**	**Foundation**	**Higher**	**Foundation**	**Higher**	**Foundation**	**Higher**
0–1,000	92%	93%	97%	98%	97%	98%	95% (3%)	96% (2%)
1,001–2,000	94%	95%	99%	99%	99%	99%	97% (2%)	98% (2%)
2,001–3,000	95%	96%	100%	100%	100%	100%	98% (2%)	98% (2%)
3,001–4,000	96%	97%	100%	100%	100%	100%	98% (2%)	99% (2%)
4,001–5,000	100%	100%	100%	100%	100%	100%	100% (0%)	100% (0%)
>5,000	100%	100%	100%	100%	100%	100%	100% (0%)	100% (0%)
No. of words	1,599 (243)	2,334 (295)	1,387 (132)	1,947 (216)	1,247 (138)	1,923 (135)	1,411 (221)	2,068 (283)
% of words from an average exam	70% (1%)	73% (0%)	71% (0%)	71% (2%)	66% (2%)	70% (2%)	69% (2%)	71% (2%)
No. of flemmas	184	229	175	209	165	228	175 (8)	222 (10)

5.3 | Number and frequency level of words used only once across four sets of exams (RQ1c)

Across the three languages, a mean of 189 flemmas was only used once across four sets of AQA exams at foundation and 249 at higher (Table 3). These flemmas represented a mean of 12% of all words in each exam at foundation and 11% at higher (Table 3). Of this set of words, a high proportion (62% at foundation and higher, averaged across the three languages) were low(er)-frequency. See Appendix S2 for additional analyses of words from flemmas used two or three times across four sets of exams.

The ordinal models revealed that the proportion of high-frequency words from this set did not vary as a function of tier, with just one exception: in the Edexcel German exams, a greater proportion of high-frequency words was used only once at foundation tier than at higher (OR: 1.28, 95% CI: [1.10–1.48], $p = .001$).

5.4 | Quantity of and changes in lexical content year-on-year (RQ2)

RQ2 investigated the number of high and low(er)-frequency flemmas and the extent to which these flemmas changed year-on-year as a function of modality and tier. For these analyses, we first calculated the number and frequency level of unique flemmas used across the four sets of exams to identify how many words an awarding organization needs to create a series of exams over time. On average, across the three languages, 1,351 unique flemmas were used in four sets of exams at foundation and 1,753 at higher, of which 56% at foundation and 55% at higher were high-frequency and, inversely, 44% at foundation and 45% at higher were low(er)-frequency (Table 4).

Using the data from RQ1a, we then computed indices of variability by dividing *the number of unique flemmas used in the body of exams* by *the sum of the number of flemmas used in each of the four individual exams*. The indices reported in Table 4 show that low(er)-frequency flemmas changed at a much higher rate than high-frequency flemmas: 45% at foundation and higher tiers for high-frequency flemmas compared with 75% at foundation and 77% at higher tier for low(er)-frequency flemmas (when averaged across the three languages). This finding was consistent across modalities, tiers, languages, and awarding organizations. See Appendix S3 for modality-specific indices.

6 | DISCUSSION

We discuss our findings with a view to addressing our overarching aim: To explore expectations about the lexical knowledge of learners in an instructed, low-exposure context by analyzing the quantities and frequency levels of words in high-stakes listening and reading exams.

6.1 | How many flemmas might students need to know to take a GCSE exam?

Our analyses showed that on average, the largest awarding organization's (AQA's) exams contained 614 flemmas at foundation and 788 at higher. In an average exam, 69% at foundation

TABLE 3 Mean (SD) cumulative coverage statistics for words from flemmas used only once across four sets of AQA exams (averaged across sets; listening and reading combined).

	French		German		Spanish		Mean	
	Foundation	**Higher**	**Foundation**	**Higher**	**Foundation**	**Higher**	**Foundation**	**Higher**
0–1,000	20% (4%)	18% (5%)	20% (6%)	17% (2%)	24% (4%)	19% (3%)	21% (5%)	18% (3%)
1,001–2,000	38% (4%)	38% (3%)	35% (5%)	34% (3%)	42% (5%)	42% (1%)	38% (5%)	38% (4%)
2,001–3,000	49% (5%)	53% (3%)	45% (6%)	45% (4%)	52% (5%)	55% (4%)	49% (6%)	51% (6%)
3,001–4,000	58% (4%)	59% (3%)	52% (6%)	53% (4%)	63% (7%)	63% (4%)	57% (7%)	58% (5%)
4,001–5,000	63% (4%)	66% (3%)	57% (5%)	58% (2%)	67% (7%)	68% (4%)	62% (6%)	64% (5%)
>5,000	100% (0%)	100% (0%)	100% (0%)	100% (0%)	100% (0%)	100% (0%)	100% (0%)	100% (0%)
No. of words	260 (12)	339 (50)	216 (27)	310 (31)	235 (19)	292 (29)	237 (26)	314 (39)
% of words from an average exam	12% (2%)	11% (1%)	11% (1%)	11% (1%)	13% (1%)	11% (1%)	12% (1%)	11% (1%)
No. of flemmas	199 (12)	253 (39)	175 (22)	249 (28)	193 (23)	246 (24)	189 (21)	249 (28)

TABLE 4 Descriptive statistics for high and low(er)-frequency words and year-on-year changes in AQA exams.

		French		German		Spanish	
		Foundation	**Higher**	**Foundation**	**Higher**	**Foundation**	**Higher**
No. of flemmas used across all years	High-frequency	802 (56%)	983 (55%)	668 (53%)	859 (50%)	805 (59%)	1043 (59%)
	Low(er)-frequency	621 (44%)	807 (45%)	590 (47%)	849 (50%)	566 (41%)	719 (41%)
	Total	1,423 (100%)	1,790 (100%)	1,258 (100%)	1,708 (100%)	1,371 (100%)	1,762 (100%)
No. of flemmas used in each year added together	High-frequency	1,714 (66%)	2,161 (67%)	1,540 (67%)	1,933 (64%)	1,734 (70%)	2,291 (72%)
	Low(er)-frequency	866 (34%)	1,087 (33%)	768 (33%)	1,090 (36%)	746 (30%)	902 (28%)
	Total	2,580 (100%)	3,248 (100%)	2,308 (100%)	3,023 (100%)	2,480 (100%)	3,193 (100%)
Variability index across years	High-frequency	47%	45%	43%	44%	46%	46%
	Low(er)-frequency	72%	74%	77%	78%	76%	80%
	Total	55%	55%	55%	57%	55%	55%

tier and 71% at higher of all running words (2,033 and 2,895, respectively) were from flemmas used consistently in *every* exam. This could suggest that for any *one* exam (in a given year), lexical expectations reflected in receptive exams are reasonable, given that the average receptive vocabulary size is estimated to be between 564 and 852 flemmas (David, 2008; Milton, 2006, 2015).

However, students must be prepared for the lexical content that could occur across a set of exams, as the lexicon of one specific year of exams cannot be predicted by teachers or students. We found that across four sets of exams, there was a mean of 1,351 unique flemmas at foundation and 1,753 flemmas at higher (averaged across three languages). Interestingly, this reflects a volume of language that is broadly comparable to the mean number of flemmas—1,283 (SD = 211) at foundation and 1,520 (SD = 185) at higher, with significant variation across languages—included on AQA's current *optional* wordlists for exams between 2015 and 2025, suggesting that the length of their current lists could be reasonable. In contrast, the equivalent wordlists produced by the second largest awarding organization (Edexcel) include many more flemmas—1,845 (SD = 71) at foundation and 2,102 (SD = 87) at higher—than the number that occurred across four exams, potentially posing an unnecessarily large teaching and learning burden.

Of potential concern is that the amount of vocabulary that *could* be encountered by a student in any one exam out of the four analyzed (1,351 or 1,753 flemmas in foundation and higher) may exceed, by some way, GCSE learners' average receptive vocabulary size, which has to date been estimated to be between 564 and 852 flemmas (David, 2008; Milton, 2006, 2015). Allaying this concern to some extent is the possibility that these vocabulary size estimates may not be valid or reliable for two reasons. First, they were based on data from small samples of GCSE learners. Second, the test instrument used (i.e., X-Lex) did not consider the curriculum when sampling test items: Dudley et al. (2024, under review) found that GCSE learners' vocabulary size estimates in French were often two or three times larger according to the Context-Aligned Two Thousand Test, where test items were sampled directly from the curriculum, relative to the X-Lex.

Finally, we found a small core of highly predictable vocabulary with very high coverage. 69% of the words used in each exam at foundation and 71% at higher were from a relatively small number of (approximately 175 at foundation and 222 at higher) flemmas that appeared in *every* exam. These words[6] seem to be extremely strong candidates for any compulsory wordlist.

In sum, our findings provide some initial (albeit somewhat mixed) indications about the expected size of the lexicon, as reflected in current optional guidance wordlists and exam content. In terms of the length of current lists, *one* awarding organization's (AQA's) list seemed broadly compatible with the amount of flemmas needed to create four exams. On the other hand, another organization's (Edexcel's) current list is substantially longer, far exceeding both assessment need and student knowledge. In terms of exam content, the number of flemmas that appeared in *every* exam was very small but provided high coverage, usefully identifying a core for any potential compulsory wordlist. Furthermore, the number of flemmas needed for unassisted comprehension of just *one* exam is broadly similar to current published estimates of vocabulary size for this proficiency of learner. However, and crucially, to be fully prepared for the lexicon in *any* exam—that is, the four sampled in this study plus *all other* exams too—students' average vocabulary size would need to be far (two to three times) larger than the current estimates that have been based on *X-lex* tests that are not aligned with the lexicon of the curriculum. Instead, vocabulary size would need to be more in line with estimations based on a curriculum-aligned test ($M = 1{,}627$, 95% CI: [1,589–1,664]), as found by

Dudley et al. (2024, under review). Estimations of vocabulary size among these learners require further investigation to further inform wordlist and exam creation.

6.2 | What is the frequency level of the words used in the GCSE exams?

A substantial proportion (86% at foundation and 87% at higher) of the words used in each exam were high-frequency. This is consistent with findings that the first 2,000 most frequent words represent at least 82% and 89% of words in written and spoken language in English, respectively (Dang & Webb, 2014; Webb & Nation, 2017).

In terms of the proportion of high-frequency words used in each exam varying as a function of modality (listening *vs.* reading), we found no systematic variability, with only a few exceptions that could be concerning. In some of these isolated cases, learners were faced, on average, with a higher proportion of low(er)-frequency words in the *listening* than the reading exams. Low(er) frequency words could represent an additional burden during listening, a skill already considered to be more difficult than reading due to (a) its ephemeral nature, which impacts learners' ability to recognize known words (van Zeeland, 2013) and infer the meaning of unknown words (van Zeeland, 2014) and (b) phonological cognateness being less transparent and prevalent than orthographic cognateness (Lubliner & Hiebert, 2011). Our findings could be used to suggest that test-developers should better ensure that listening exams represent the tendency for spoken language to contain more high-frequency flemmas than written language.

In terms of frequency distributions across foundation and higher tier papers, another potential concern could be that there were two cases (AQA's Spanish and Edexcel's French exams) where foundation exams had, perhaps counterintuitively, a greater proportion of low (er)-frequency words than higher exams. This may be in part due to the use of highly topic-specific—and by definition low(er) frequency—vocabulary in the exams, which may be *proportionally* larger in some foundation exams due to them being shorter than higher exams.

However, such subtle distinctions in frequency distributions across modality (listening and reading) and tier (foundation and higher) may not be meaningful or essential given the relatively low proficiency of GCSE learners and their very curriculum-bound exposure to the language. We suggest that perhaps simply having a *similar* proportion of high-frequency words across these parameters might be a default expectation and feasible to implement.

We identified a very small set of mainly high-frequency words that were highly predictable across all exams, as noted above. However, the low predictability of the *other* words poses a challenge. A strong indication of this unpredictability is that just over one in 10 (12% at foundation and 11% at higher) of all words used in each exam were from flemmas that had only ever appeared *once* across four exams, and almost two thirds (62% at both foundation and higher) of these "single-use" words were low(er)-frequency. Moreover, low(er)-frequency flemmas consistently changed at a much higher rate between exams than high-frequency flemmas, regardless of modality, tier, or language. Although expected to some extent, the *magnitude* of this difference in rates of change further suggests that a non-negligible proportion of the GCSE lexicon was both unpredictable and low(er)-frequency.

Arguably, the use of low(er)-frequency words may not always be problematic if these words are cognates. For instance, Lindgren and Muñoz (2013) found that learners achieved higher listening and reading scores during early foreign language learning when the cognate linguistic distance (i.e., lexical similarity) was smaller between the L1 and the target English. However,

our posthoc analyses revealed that only a tiny percentage (4% at foundation; 3% at higher) of the low(er)-frequency words were orthographic cognates in an average AQA exam.[7] This suggests that most of the low(er)-frequency words included in an average exam were *not*, in fact, easily identifiable to English-speaking beginner-to-low-intermediate learners.

For virtually all L2 readers and listeners, there will almost always be *some* unknown vocabulary in any text. However, extensive, uncontrolled, and rapidly changing use of low(er)-frequency words can increase the proportion of unknown vocabulary and, in turn, be problematic for test-takers for several reasons. First, low(er)-frequency words are generally less likely to be known than high-frequency words (David, 2008; Milton, 2006, 2015), simply because the chances of having encountered them are smaller. Second, teachers and learners are less able to predict *which* low(er)-frequency words will be used in any future exam, as the past exams used for revision are unlikely to include them. Of course, some unpredictability is expected and necessary, but for learners with just 400–450 hours of instruction and little exposure outside school, our findings suggest there is room to make the lexical content a little more predictable.

Another reason for concern is that low(er)-frequency words place a greater reliance on inferencing. Given that 95% coverage is necessary for good comprehension (e.g., Schmitt et al., 2011) and that learners are less likely to know low(er)-frequency words, it is likely that the current exams are substantially tapping into inferencing skills. Test writers may be using inferencing skills to differentiate between abilities. After all, lexical inferencing is highly variable among L2 learners, especially in listening (van Zeeland, 2014), and develops as a function of word knowledge (e.g., Hatami & Tavakoli, 2013). However, lexical inferencing may have been used as a principal and widespread differentiator without test designers explicitly stating whether any systematic attention is given to the vocabulary used around the less familiar or unfamiliar words (i.e., low(er) frequency words or words not on the guidance wordlists). Use of lexical inferencing as a key differentiator could be problematic given that, for example, (a) inferencing use and success is highly variable even among first-language speakers (Wesche & Paribakht, 2009) and (b) inferencing was not stated as a major objective in the Department for Education's (2015) subject content.

6.3 | Potential implications for policy-makers, curriculum-designers, and test-developers

It has been argued that high-quality L2 assessments should be "rooted in a principled and verifiable body of content, coming from a lesson in a textbook, a syllabus, standards, or a model of L2 proficiency" (Bachman & Palmer, 2010; Purpura, 2016, p. 191). Our findings, however, suggest that 16-year-olds, after about 400–450 hours of instruction, have not to date been tested on a lexicon that is "appropriate to this level" or to the general communicative purpose of the assessments (Department for Education, 2015, p. 4, a document which applies until 2025).

Having a largely pre-defined wordlist could usefully washback into school curricula and pedagogy and thus help test-developers, teachers, and test-takers to have a clearer idea about the language needed to progress through these early stages of classroom learning (Dang et al., 2020). But how big should such lists be? Our findings suggest that lists of *approximately* 1,350 lexical items at foundation tier and 1,750 at higher may be appropriate for several reasons. First, awarding organizations must be able to sample knowledge of a defined body of language (akin to "achievement") across the years, a process they are required and regulated by

Ofqual (2021b) to monitor every three-to-five years, while also allowing learners to demonstrate their use of knowledge in communicative and meaningful contexts (akin to "proficiency"). Our analyses showed that 1,351 and 1,753 flemmas allow the creation of four exams (at foundation and higher tiers, respectively) and thus provide sufficient scope to create meaningful opportunities for target language use. Therefore, assuming the amount of text in exams remains similar in the future, wordlists of roughly these lengths would achieve desirable coverage of *any* single exam to align with findings that at least 95% coverage is needed for unassisted comprehension testable by *any* comprehension question (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). (Further research, however, is needed to investigate the level of coverage required to answer the specific comprehension questions included in the exam papers.) Second, evidence suggests that schoolchildren in England learn approximately 3.8–4.3 (mean 4.1) words per hour (Milton & Meara, 1998). Given that they have two hours per week of instruction in the first three years (Years 7–9; aged 11–14) of secondary (high) school and two and a half hours in the GCSE years (Years 10 and 11; aged 14–16) and spend 174 weeks in language lessons during secondary school,[8] an approximate 1,500 words seems a reasonable target, when averaged across foundation and higher tier students. Finally, these numbers also happen to broadly align with a recommendation—arrived at by entirely different means—made by Milton (2015), who proposed that learning approximately 1500 words could be a reasonable target for GCSE learners.

We recommend that high-frequency words are prioritized when developing these lists, as they cover a high proportion of the words in exams, as in life, and thus maximize students' return for their learning effort (Laufer & Nation, 2012; Nation, 2013). As Dang et al. (2020, p. 19) highlight, "[g]iven the importance of high-frequency words and [...] learners' insufficient knowledge of these words, it is essential for teachers and course designers to ensure that learners have mastered these words." Our arguments thus align with many others who have advocated for the use of wordlists informed by frequency data drawn from large corpora of written and spoken texts (Brezina & Gablasova, 2015; Browne, 2014; Cobb & Laufer, 2021; Nation, 2016; Schmitt, 2010). Because corpora created specifically for beginner-to-low-intermediate learners do not exist and the GCSE exams are, by definition, for *general* purposes, corpora derived from a large number and wide range of (con)texts are likely to be crucial for awarding organizations when developing the wordlists, at least until further research can be done.

So, what proportion of flemmas on any wordlist should be high-frequency? The pool of high-frequency words must be sufficiently large enough for awarding organizations to sample, year-on-year, without being too predictable. To create four exams, the largest awarding organization drew on a total of 1,351 flemmas at foundation and 1,753 at higher, of which only 56% at foundation and 55% at higher tier were high-frequency. Nevertheless, these high-frequency flemmas and their members gave rise to a relatively high proportion (86% at foundation and 87% at higher) of *running words* in an average exam being high-frequency. To date, however, a substantial proportion (78% at foundation and 80% at higher) of the high-frequency words used in *every* exam have come from a *very* small set of high-frequency flemmas —just 163 at foundation and 208 at higher. This small set has comprised nearly a quarter (22% at foundation and higher) of all high-frequency flemmas used at least once across the corpora of exams and includes mainly function words and core verbs (e.g., *be*, *have*, *go*, *do*, and *make*). It therefore seems desirable for the awarding organizations to draw on a *wider range* of high-frequency words, thus increasing the pool of widely used flemmas from which to sample year-on-year. It could therefore be desirable to stipulate that approximately 85% of the lexical items

on any wordlist are high-frequency. This would mean that some high-frequency flemmas (function words and high-frequency verbs) will inevitably be re-used every year, whereas other high-frequency words will often—but not *always*—be re-used in exams, thus resulting in more varied sampling from a wider set of words that are known to be used most often across contexts. Critically, we do not suggest that low(er)-frequency words should be *avoided*. We instead propose that low(er)-frequency words be used with greater moderation than hitherto and that explicit decisions are made about *which* mid-to-low-frequency words serve the likely needs of these learners.

Finally, our findings suggest that lexical inferencing skills could be unintentionally contributing to assessment outcomes. Some continued testing of lexical inferencing is clearly desirable, but with constraints and greater systematicity across awarding organizations, languages, and years. To this end, an explicit statement on the proportion of words in listening and reading texts that can be *off* any pre-defined wordlist could be beneficial, combined with an explicit mention of lexical inferencing as part of the subject content.

Note that at the time of finalizing this academic article, many of these recommendations had been operationalized ("as is" or adapted) in the revised GCSE subject content (Department for Education, 2022) and in the forthcoming wordlists, specifications, and sample assessment material from the awarding organizations.

7 | CONCLUSION

This article has provided insights into the lexical content of high-stakes, national exams aimed at beginner-to-low-intermediate learners of French, German, and Spanish. Key findings include: (a) nearly half of the flemmas that have been used, over four exams, were low(er)-frequency; (b) on average, about one in 10 words have only ever been used in one exam; and (c) about 1,350 and 1,750 flemmas were used to create four sets of exams. This number of flemmas exceeds current receptive vocabulary size estimates for 16-year-olds. However, those estimates were based on tests that were not aligned with the curriculum content and more accurate estimates are needed (see Dudley et al., 2024). Overall, it seems likely that many students have had to rely heavily on inferencing to understand the texts.

At the same time, we found a very small core of highly consistent language (175 flemmas at foundation and 222 at higher) was used in *every* exam and covered a very large proportion of the total words used (69% at foundation and 71% at higher). It could therefore be desirable to stipulate that a high proportion of any compulsory wordlist should consist of high-frequency words so as to both (a) keep a greater number of generally useful words constant year-on-year and (b) increase the pool of generally useful words that could change year-on-year.

We emphasize that word frequency alone cannot be used to comment on the difficulty of exams. Clearly, other factors, such as "semantic neutrality, length, part of speech, polysemy, morphological regularity, cognateness, [and] orthographic transparency" (Hashimoto, 2021, p. 182), also influence lexical difficulty, and the difficulty of any written or spoken text involves many dimensions beyond the lexicon. Similarly, objective word frequency information (e.g., corpus-based data and lexical coverage) should not be the sole criterion for wordlist development, which can helpfully also draw on subjective criteria (e.g., teacher evaluations; Dang et al., 2020; He & Godfroid, 2019; Marsden et al., 2023).

Assuming that it is not possible to substantially increase curriculum time given to French, German, or Spanish, then more clearly defining the lexical content for GCSE exams, with a

higher concentration of high-frequency words, could help tighten the link between "what is taught" and "what can be assessed" and allow students to get more return for their learning effort (Laufer & Nation, 2012; Nation, 2013). While reducing the likely current burden on learners' use of repair strategies, such as lexical inferencing, a compulsory wordlist for exam creation does not have to eradicate the testing of lexical inferencing, as a small defined proportion of "off list" words could still be tested. Moreover, it is extremely unlikely that test-takers will know every word in an exam, especially given vocabulary size estimates for this population of learners. Thus, some lexical inferencing would still be required.

Clearly defining lexical content could help teachers, curriculum designers, and textbook publishers in low-exposure instructed contexts to prioritize useful language. And perhaps equally crucially, a defined core lexical content could reveal to teachers and learners the scope for and amount of time available for developing *personalized* lexicons to reflect individuals' interests. Also worthy of future investigation is whether a more clearly defined lexical content could contribute to leveling the playing field between socio-economically disadvantaged students and their more privileged peers who may have more opportunities for engagement with the target language (e.g., on vacations).

AUTHOR CONTRIBUTIONS

Amber Dudley: Conceptualization (supporting); methodology (lead); formal analysis (lead); investigation (lead); resources (supporting); data curation (lead); writing—original draft (lead); writing—review and editing (lead); visualization (lead); project administration (lead). **Emma Marsden**: Conceptualization (lead); methodology (supporting); formal analysis (supporting); investigation (supporting); resources (lead); writing—original draft (supporting); writing—review and editing (lead); supervision (lead); project administration (lead); funding acquisition (lead).

ACKNOWLEDGMENTS

This research was supported by funding from the Department for Education for England awarded to the former National Centre for Excellence for Language Pedagogy (2018–2023) and to Professor Emma Marsden at the University of York (2023–2024) and by funding from Research England, the Higher Education Innovation Funding, Economic and Social Research Council Impact Acceleration Account, and the University of York. The authors would like to thank colleagues at the former National Centre for Excellence for Language Pedagogy (now Language-Driven Pedagogy) for preparing the exam papers for profiling; Dr Giulia Bovolenta for running the Python script to split the German compounds; and to Dr Natalie Finlayson and Professor Laurence Anthony for their work on the development of the Multiling Profiler (www.multilingprofiler.net).

OPEN RESEARCH BADGES

This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at https://osf.io/kmn39/?view_only=529ccb57aa124c30b696dc1ee829afe6

ORCID

Amber Dudley http://orcid.org/0000-0003-2904-9150

ENDNOTES

[1] GCSEs are part of the national curriculum in England, which outlines the programs of study and attainment targets for all subjects in primary and secondary education.

[2] AQA held 78% of the GCSE French, German, and Spanish market in the 2021/2022 academic year (Ofqual, 2023).

[3] Software is available for French via Tom Cobb's Lextutor website (http://www.lextutor.ca) but has never been used by awarding organizations, as far as we are aware.

[4] In 2022, the requirement to test words off the optional wordlists was lifted and awarding organizations have been allowed to gloss any words that do not appear in their specification's wordlist (Ofqual, 2021a). Initially, this measure was introduced to acknowledge disruption caused by the pandemic but will now remain in anticipation of the revised GCSE qualifications in French, German, and Spanish (Ofqual, 2022), informed by some of the current findings.

[5] Adding modality as a predictor would have required us to investigate the words, for RQ1b, from flemmas used in both listening *and* reading (but not in only one or the other) or used within the same sets of exams; and, for RQ1c, from flemmas only ever used in a listening *or* reading exam but never both within the same set of exams. Such analyses would have introduced a level of granularity beyond the scope of our study.

[6] These words are available on our OSF repository and have been shared among networks of teachers.

[7] See Appendix S4 for information about how orthographic cognates were identified.

[8] The 174 weeks in language lessons during secondary school equates to 36 weeks a year for five years, allowing for tests, extracurricular events, and half a term in Year 11 when students take their GCSEs.

REFERENCES

ACTFL. (n.d.). *Assigning CEFR ratings to ACTFL assessments*. https://www.actfl.org/uploads/files/general/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf

College voor Toetsen en Examens. (n.d.). *Over het Staatsexamen Nt2: Programma I en II*. https://www.staatsexamensnt2.nl/item/programma-i-en-ii

Adolphs, S. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438. https://doi.org/10.1093/applin/24.4.425

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

AQA. (2023). *AQA | GCSE | French | Assessment resources*. https://www.aqa.org.uk/subjects/languages/gcse/french-8658/assessment-resources

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*(1), 1–22. https://doi.org/10.1093/applin/amt018

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

Browne, C. (2014). A New General Service List: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction*, *3*(2), 1–10. https://doi.org/10.7820/vli.v03.2.browne

Christensen, R. H. B. (2019). *Ordinal: Regression models for ordinal data*. https://cran.r-project.org/package=ordinal.

Churchward, D. (2019). *Recent trends in modern foreign language exam entries in anglophone countries*. Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/844128/Recent_trends_in_modern_foreign_language_exam_entries_in_anglophone_countries_-_FINAL65573.pdf

Cobb, T., & Horst, M. (2004). Is there room for an academic word list in French? In P. Bogaards, & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 15–38). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.10.04cob

Cobb, T., & Laufer, B. (2021). The nuclear word family list: A list of the most frequent family members, including base and affixed words. *Language learning*, *71*(3), 834–871. https://doi.org/10.1111/lang.12452

Coffey, S. (2016). Choosing to study modern foreign languages: Discourses of value as forms of cultural capital. *Applied Linguistics*, *39*(4), 462–480. https://doi.org/10.1093/applin/amw019

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238. https://doi.org/10.2307/3587951

Curcin, M., & Black, B. (2019). *Investigating standards in GCSE French, German and Spanish through the lens of the CEFR*. Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/844034/Investigating_standards_in_GCSE_French__German_and_Spanish_through_the_lens_of_the_CEFR.pdf

Dang, T. N. Y., & 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

Dang, T. N. Y., Webb, S., & Coxhead, A. (2020). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, *26*(4), 617–641. https://doi.org/10.1177/1362168820911189

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

Davies, M., & Davies, K. H. (2017). *A frequency dictionary of Spanish*. Routledge. https://doi.org/10.4324/9781315542638

DeBruine, L., Krystalli, A., & Heiss, A. (2021). *faux: Simulative for factorial designs*. https://cran.r-project.org/web/packages/faux/

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. (2021a). *GCSE MFL subject content review*. https://consult.education.gov.uk/ebacc-and-arts-and-humanities-team/gcse-mfl-subject-content-review/

Department for Education. (2021b). *GCSE MFL subject content review: supporting document*. https://consult.education.gov.uk/ebacc-and-arts-and-humanities-team/gcse-mfl-subject-content-review/supporting_documents/GCSE%20MFL%20subject%20content%20document.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

Department for Education. (2023). *Key stage 4 performance*. https://explore-education-statistics.service.gov.uk/data-tables/permalink/c04c8db0-0cd3-4dd7-fb3a-08dbf5f3097e

Dudley, A., Marsden, E., & Bovolenta, G. (2024). The Context-Aligned Two Thousand Test: A new test of French high-frequency vocabulary size for beginner-to-low intermediate proficiency adolescent learners. OSF Preprints. https://doi.org/10.31219/osf.io/x6bzs

Education Bureau. (2021, October). *Preamble to the development of the wordlists for the English language curriculum*. https://www.edb.gov.hk/en/curriculum-development/kla/eng-edu/references-resources/Wordlists_preamble.html

éduscol. (2020). *Liste de fréquence lexicale*. https://eduscol.education.fr/186/liste-de-frequence-lexicale

Finlayson, N., Marsden, E., & Anthony, L. (2022). MultilingProfiler (Version 3) [Computer software]. University of York. Accessed 12/01/2022 at https://www.multilingprofiler.net/

Finlayson, N., Marsden, E., & Anthony, L. (2023). Introducing MultilingProfiler: An adaptable tool for analysing the vocabulary in French, German, and Spanish texts. *System*, *118*, 103122. https://doi.org/10.1016/j.system.2023.103122

van Goethem, K., & Amiot, D. (2019). Compounds and multi-word expressions in French, *Complex lexical units* (pp. 127–152). De Gruyter. https://doi.org/10.1515/9783110632446-005

Haastrup, K. (1991). *Lexical inferencing procedures or talking about words: Receptive procedures in foreign language learning with special reference to English*. Gunter Narr.

Hashimoto, B. J. (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

Hatami, S., & Tavakoli, M. (2013). The role of depth versus breadth of vocabulary knowledge in success and ease in L2 lexical inferencing. *TESL Canada Journal*, *30*(1), 1. https://doi.org/10.18806/tesl.v30i1.1123

He, Q., & Black, B. (2019). *Statistical evidence pertaining to the claim of grading severity in GCSE French, German and Spanish and the impact of statistical alignment of standards on outcomes*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/844033/Statistical_Evidence_Report_-_ISC_-_FINAL65574.pdf

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, P. (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

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kremmel, B. (2021). Selling the (word) family silver?: A response to Webb's "lemma dilemma". *Studies in Second Language Acquisition*, *43*(5), 962–964. https://doi.org/10.1017/S0272263121000693

Kremmel, B., Indrarathne, B., Kormos, J., & Suzuki, S. (2023). Unknown vocabulary density and reading comprehension: Replicating Hu and Nation (2000). *Language Learning*, *73*(4), 1127–1163. https://doi.org/10.1111/lang.12622

Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*, *54*(4), 1076–1085. https://doi.org/10.1002/tesq.3004

Laufer, B. (2023). 10 From research to a national curriculum: The case of a lexical syllabus. In G. Erickson, C. Bardel & D. Little (Eds.), *Collaborative research in language education* (pp. 151–164). De Gruyter. https://doi.org/10.1515/9783110787719-011

Laufer, B., & Nation, P. (2012). Vocabulary. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 163–176). Routledge.

Laufer, B., & Ravenhorst-Kalovski, G. C. (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

Lenth, R. (2021). *emmeans: Estimated marginal means, aka least squares means (v.1.7.1-1)*. https://cran.r-project.org/web/packages/emmeans/index.html

Lindgren, E., & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. *International Journal of Multilingualism*, *10*(1), 105–129. https://doi.org/10.1080/14790718.2012.679275

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

Lubliner, S., & Hiebert, E. H. (2011). An analysis of English–Spanish cognates as a source of general academic language. *Bilingual Research Journal*, *34*(1), 76–93. https://doi.org/10.1080/15235882.2011.568589

Lüdecke, D. (2021). *sjPlot: Data visualization for statistics in social science. R Package Version 2.8.10*. https://cran.r-project.org/web/packages/sjPlot/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

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

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* (Vol. 45, 1st ed.). Multilingual Matters.

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

Milton, J., & Alexiou, T. (2009). Vocabulary size and the common European framework of reference for languages. In B. Richards, M. Daller, D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 194–211). Palgrave Macmillan.

Milton, J., & Meara, P. (1998). Are the British really bad at learning foreign languages. *The Language Learning Journal*, *18*(1), 68–76. https://doi.org/10.1080/09571739885200291

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

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

Nation, P. (2012). *The BNC/COCA word family lists*. https://people.wgtn.ac.nz/paul.nation

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

Nation, P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company. https://doi.org/10.1075/z.208.

Ofqual. (2016). *GCSE subject level guidance for modern foreign languages*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/517346/gcse-subject-level-guidance-for-modern-foreign-languages.pdf

Ofqual. (2021a). *Decisions on arrangements for non-exam assessment and fieldwork requirements for students entering qualifications in 2022*. https://www.gov.uk/government/consultations/arrangements-for-non-exam-assessment-for-qualifications-in-2022/outcome/decisions-on-arrangements-for-non-exam-assessment-and-fieldwork-requirements-for-students-entering-qualifications-in-2022

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

Ofqual. (2022). *Covering requirements for the assessment of vocabulary*. https://www.gov.uk/government/consultations/proposed-changes-to-the-assessment-of-modern-foreign-language-gcses-from-2023/covering-requirements-for-the-assessment-of-vocabulary

Ofqual. (2023). *Data tables for annual qualifications market report: academic year 2021 to 2022*. https://assets.publishing.service.gov.uk/media/652d42766b6fbf0014b756ff/AQMR_2021_to_2022_academic_year.ods

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. *8th Language resources and evaluation conference*.

Parrish, A., & Lanvers, U. (2019). Student motivation, school policy choices and modern language study in England. *The Language Learning Journal*, *47*(3), 281–298. https://doi.org/10.1080/09571736.2018.1508305

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

Pearson Edexcel. (2023). *Pearson Edexcel GCSE French (2016) | Pearson qualifications | Exam materials*. https://qualifications.pearson.com/en/qualifications/edexcel-gcses/french-2016.coursematerials.html#filterQuery=Pearson-UK:Category%2FExam-materials

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal*, *100*(S1), 190–208. https://doi.org/10.1111/modl.12308

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

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, *7*(2), 105–126.

Schmitt, N. (2010). *Researching vocabulary*. Palgrave Macmillan UK. https://doi.org/10.1057/9780230293977

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

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/S0261444812000018

Sternberg, R. J. (1987). Most vocabulary is learned from context. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 89–105). Erlbaum.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*(4), 577–607. https://doi.org/10.1017/S0272263109990039

Stratton, T., & Zanini, N. (2018). Evaluating the impact of the introduction of reformed GCSE MFL assessments in 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/844031/Evaluating_the_impact_of_the_introduction_of_reformed_GCSE_MFL_assessments_in_2018_-_FINAL65572.pdf

Taylor, F., & Marsden, E. J. (2014). Perceptions, attitudes, and choosing to study foreign languages in England: An experimental intervention. *The Modern Language Journal*, *98*(4), 902–920. https://doi.org/10.1111/modl.12146

The Education (National Curriculum) (Modern Foreign Languages) Order 1991, Pub. L. No. 2567 (1991).

Thornbury, S. (2002). *How to teach vocabulary*. Longman.

Tschirner, E., & Möhring, J. (2019). *A frequency dictionary of German*. Routledge. https://doi.org/10.4324/9781315620008

Tuggener, D. (2016). *Incremental coreference resolution for German*. [PhD Thesis]. University of Zurich.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1), 79–95. https://doi.org/10.1017/S0272263108080042

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.

Wesche, M. B., & Paribakht, T. S. (2009). *Lexical inferencing in a first and second language: Cross-linguistic dimensions*. Multilingual Matters.

West, M. (1953). *A general service list of English words*. Longman Green.

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of-school exposure: How do word-related variables and proficiency influence receptive vocabulary learning? *Language learning*, *70*(2), 349–381. https://doi.org/10.1111/lang.12380

van Zeeland, H. (2013). L2 vocabulary knowledge in and out of context. *Australian Review of Applied Linguistics*, *36*(1), 52–70. https://doi.org/10.1075/aral.36.1.03van

van Zeeland, H. (2014). Lexical inferencing in first and second language listening. *The Modern Language Journal*, *98*(4), 1006–1021. https://doi.org/10.1111/modl.12152

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479. https://doi.org/10.1093/applin/ams074

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

Zipf, G. (1935). *The psychobiology of language: An introduction to dynamic philology*. MIT Press.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Dudley, A., Marsden, E. (2024). The lexical content of high-stakes national exams in French, German, and Spanish in England. *Foreign Language Annals*, 1–28. https://doi.org/10.1111/flan.12751