
Article:
Marsden, Emma orcid.org/0000-0003-4086-5765, Dudley, Amber orcid.org/0000-0003-2904-9150 and Hawkes, Rachel (2023) Use of word lists in a high-stakes, low-exposure context:Topic-driven or frequency-informed. The Modern Language Journal. 669–692. ISSN 1540-4781

https://doi.org/10.1111/modl.12866

DOI: 10.1111/modl.12866

The Modern Language Journal

ORIGINAL ARTICLE

Use of word lists in a high-stakes, low-exposure context: Topic-driven or frequency-informed

Emma Marsden[1] | **Amber Dudley**[1] | **Rachel Hawkes**[2]

[1]Department of Education, University of York, York, UK

[2]The Cam Academy Trust, International Education and Research, Cambridge, UK

Correspondence
Amber Dudley, Department of Education, University of York, York, YO10 5DD, UK.
Email: amber.dudley@york.ac.uk

Funding information
University of York; Department for Education, UK Government; Economic and Social Research Council's Impact Acceleration Account Higher Education Innovation Fund Research England

Abstract

The awarding organizations that create and administer high-stakes assessments for beginner-to-low-intermediate 16-year-old learners of French, German, and Spanish in England provide optional topic-driven word lists as guides for teachers and textbook writers. Given that these lists are developed by the awarding organizations, they exert a powerful washback effect on teaching and learning. However, we do not know how much of these lists have actually been used in exams. We therefore analyzed the extent to which these lists have been used when developing the General Certificate of Secondary Education listening and reading exams, a corpus totaling 116,647 words. One key finding showed that approximately half of the awarding organizations' lists had never been used in any of the exams to date. Given recent changes to curriculum policy, we also investigated how word list type—frequency-informed versus the awarding organizations' topic-driven lists—affected lexical coverage of the exams. Overall, our findings suggested that using the topic-driven lists was likely to be a suboptimal use of lesson time, as they did not provide learners with enough words to understand any given text with ease. Frequency-informed word lists, however, seemed to better prepare learners for the exams.

KEYWORDS
foreign language education, high-frequency vocabulary, high-stakes testing, lexical coverage, lexical selection, topic-driven vocabulary

Core decisions when designing a language curriculum relate to whether the approach is more synthetic, whereby language content is predefined and sequenced, or more analytic, whereby the curriculum is structured around different tasks or topics, for example, rather than the language itself (Ellis, 2019; Wilkins, 1976). Where a curriculum sits on this analytic–synthetic continuum is, in part, determined by the existence or otherwise of a word list, the strength of the association between the word list and the assessments, and the principles underlying word selection. The analysis and discussion provided in this article aim to inform this debate, by focusing on changes to foreign language (French, German, and Spanish) education in England.

Since the 1980s, foreign language education in England has adopted a largely topic-driven curriculum, in part influenced by (mis- or over-)interpretations of communicative approaches to language teaching. Language educators, textbook publishers, and awarding organizations have thus selected vocabulary based on topic-specific sets of words, such as the weather, food and drink, free-time activities, daily routine, and travel. This approach to lexical selection is most clearly evidenced in the topic-driven word lists published by the three commercial awarding organizations—the Assessment and Qualifications Alliance (AQA, 2016), the English branch of the Welsh Joint Education Committee (Eduqas, 2019), and Pearson Edexcel (2018)—as part of their specifications for the General Certificate in Secondary Education (GCSE) exams taken by approximately 250,000 16-year-olds every year in England (Churchward, 2019).

Awarding organizations produce these lists as guides for teachers to plan schemes of work and for publishers to write textbooks. As such, they have not been required to use their lists when developing GCSE exams. These lists, however, exert powerful washback effects on what is taught, as can be seen, for example, in the high-selling textbooks aimed at GCSE students (Hawkes & Lillington, 2016a, 2016b, 2016c, 2016d). These effects are perhaps magnified by the fact that curriculum time is limited, with students only receiving between 400 and 450 hours of instruction before taking their GCSE exams. Most of this instruction takes place during secondary school given the limited provision of foreign language education in primary schools and, generally, a lack of continuity between primary and secondary education (Graham et al., 2016). Thus, valuable curriculum time is spent teaching and learning words from the lists. Yet, the strength of the association between these lists and the words included in the exams is unknown. It may be that a strong dissociation between what is learned and what is assessed is contributing to the perceived (Coffey, 2016) and actual (Curcin & Black, 2019) difficulty of foreign languages at GCSE, given that students feel more motivated to learn when they (feel like they) are achieving more (Garon–Carrier et al., 2016).

In November 2019, in an attempt to understand and potentially address the challenges posed by the current curriculum and exams, the Department for Education in England requested a review of the GCSE subject content (henceforth, curriculum). One aim of that review was to consider the benefits of introducing a word list that would be informed by word frequency and used to develop future exams (Department for Education, 2021). The review culminated in January 2022, with the Department for Education (2022a) announcing a revised GCSE curriculum for French, German, and Spanish for first examination in 2026. This new curriculum stipulates that at least 85% of the items on the awarding organizations' word lists must be high frequency at both foundation and higher tier. For some subjects, such as foreign languages, teachers must enter their students for either foundation or higher tier: Foundation tier allows students to achieve Levels (grade) 1−5, and higher tier Levels 4−9.

Research on the value of using frequency-informed word lists as preparation for high-stakes exams, however, is limited (Dang et al., 2020). It is therefore of interest for both research and practice to know whether using a new frequency-informed list might serve as a comparable, or perhaps better, preparation for the current GCSE exams, relative to the awarding organizations' topic-driven lists. To define high frequency, views differ as to whether the first 2,000 (Nation, 1990, 2001; Read, 2000; Schmitt, 2000; Thornbury, 2002) or 3,000 (Schmitt & Schmitt, 2014) most frequent words should be used as the cut-off point. The 3,000 view is based, in part, on findings estimating that between 2,000 and 3,000 words are needed to understand a spoken conversation (Adolphs & Schmitt, 2003; van Zeeland & Schmitt, 2013). For the current analyses and the new GCSE curriculum, however, we

define high frequency as the first 2,000 most frequent words for several reasons, including our focus on the initial stages of learning, the vocabulary size of GCSE learners (averaging some way below 1,000 words; David, 2008; Milton, 2006, 2015), and the limited classroom time available.

The present study fed into the later part of the review and consultation process and set out to explore (a) the extent to which awarding organizations have used their lists when creating exams, and (b) the extent to which a frequency-informed list might prepare students for texts that are deemed to be broadly appropriate for their age and proficiency—that is, the GCSE French, German, and Spanish listening and reading exams—relative to the current lists that the awarding organizations provide.

BACKGROUND LITERATURE

We first outline recent developments in foreign language education in England, as this informed the purpose and nature of our analyses. We then consider existing research in four areas relevant to our analyses: the use of word lists for language curricula and assessment, the importance of word frequency for coverage and comprehension, the implications of coverage for lexical inferencing, and the role of semantic relatedness in the early stages of learning.

Modern foreign language education in England

The perceived and evidenced difficulty of the GCSE exams

In September 2004, studying a foreign language at GCSE became optional. Since then, there has been a sharp decline of almost 50% in the number of students taking GCSE exams in French and German in England, countered, to some extent, by a twofold increase in the number of students studying Spanish (Churchward, 2019). One of the many reasons for these declines is thought to be the perception that foreign languages are more challenging than other school subjects (Coffey, 2016; Graham et al., 2016; Parrish & Lanvers, 2019; Taylor & Marsden, 2014). Indeed, He and Black (2019), on behalf of the Office of Qualifications and Exams Regulation (Ofqual, the government's regulatory body), found that GCSE French and German—but not Spanish—were systematically more difficult, in statistical terms, than many other GCSE subjects. In response, Ofqual (2019) adjusted grading standards in GCSE French and German. Although such adjustments can address disparities in standards between languages and other subjects, they cannot tell us about the causes of the perceived and/or actual difficulty of foreign languages.

One possible factor contributing to the difficulty of foreign languages could indeed be the lexical content of the listening and reading exams. Webb and Paribakht (2015), for example, analyzed the coverage (i.e., percentage of headwords and family members) that the Academic Word List (Coxhead, 2000) provided of English language listening and reading proficiency tests used for admission purposes in Canadian universities. They reported that the lexical content of these tests did not reflect the vocabulary that learners would encounter in their studies, given that the Academic Word List provided much lower coverage of these high-stakes tests than of academic texts.

Similarly, Jin et al. (2016) explored the coverage that the official vocabulary list (of 1,507 entries) for the English language curriculum in China provided of 859 reading texts from high school entrance exams. Their study found that this list, on average, covered approximately 93% of the words in the reading texts. Given that knowledge of at least 95% of the words is typically needed to understand written texts (Schmitt et al., 2011), they concluded that there was a mismatch between what is taught and what is assessed.

Very little research, however, has examined the specific lexical demands of the GCSE exams. In our recent analysis (Dudley & Marsden, 2023), we observed extensive use of unpredictable, low(er)

frequency, and likely unknown words in four sets of GCSE exams and suggested that this may be contributing to the perceived and actual difficulty of foreign languages.

A similar study (Stratton & Zanini, 2018) found that the 2015 reforms to the GCSE curriculum (first examined in 2018) resulted in an unexpected and, critically, undesirable increase in the difficulty of the exams "due to an increase in the demand of the vocabulary used in the reading and listening texts" (p. 5), as assessed by subject experts. Although included in the statistical models, lexical familiarity (i.e., the proportion of words in the exams taken directly from the word lists) did not moderate this increase in difficulty. One limitation of this study, however, was that it only focused on two sets of exams: one developed before the reforms and another after. Our study extends this line of enquiry by analyzing the coverage that the awarding organizations' current (or other frequency-informed) word lists provide for four sets of GCSE exams developed under the current GCSE curriculum. It could indeed be this very relationship between the word lists and the exams that—at least in part—adversely affects the achievement–motivation cycle.

The achievement–motivation cycle has been widely documented in both the general (Vu et al., 2022) and language (Erler & Macaro, 2011; Graham et al., 2016; Taylor & Marsden, 2014) education literature. In general, these studies have found that students are more likely to feel motivated and continue studying languages at GCSE and beyond when they have a positive perception of their ability to learn effectively and make progress in the language. It is therefore possible that any current misalignment between what is tested and what is taught in language lessons may be having a detrimental effect on students' perception of their language-learning abilities and, consequently, their motivation to continue studying languages at more advanced levels.

At present, the extent to which the awarding organizations have used their word lists in the GCSE exams across the years is unknown. We therefore do not know whether teaching vocabulary from these lists is an effective use of curriculum time in terms of preparing students for the exams. Attesting to this being a possible problem, Pearson Edexcel (2020) received feedback (sought around a similar time to the GCSE review process) from 400 students from six schools that "spoken extracts [and texts] contain[ed] too many words not on vocabulary lists" (p. 11). In response, the organization committed to including "fewer non-vocabulary list words" (p. 11) in their exams.

In sum, it may be that the lexical content of the exams and its relationship with the awarding organizations' current word lists have contributed, at least in part, to the perceived difficulty of the subject and the low uptake of languages at GCSE and beyond. Thus, the current study set out to explore the extent to which the awarding organizations' topic-driven lists align with the vocabulary used in the GCSE exams, relative to frequency-informed lists. Specifically, we focused on the GCSE listening and reading exams (see Online Supporting Information A for a description) for several reasons, including the ease and reliability of analyzing receptive assessments relative to production data.

Awarding organizations' word lists: current and future

The current GCSE curriculum (Department for Education, 2015) does not stipulate the volume and/or frequency of the vocabulary that the exams should test, nor how much of the word lists test writers should use, or how often, when creating the exams. As such, it has been at the awarding organizations' discretion as to whether they publish word lists as part of their specifications. Although all three do (AQA, 2016; Eduqas, 2019; Pearson Edexcel, 2018), the lists are nonexhaustive and serve only as guides for schemes of work and textbooks. In fact, until very recently , the awarding organizations were required by Ofqual (2021a) to include words *not* on their word lists. This requirement changed only very recently in a response to the impact of the pandemic. As such, there has been no way of knowing how the lexical content of the GCSE exams has been selected.

A further problem in establishing the relations between the lists and the exam content is that a substantial percentage of the lists are made up of multiword phrases (e.g., "ça dépend [it depends]" or other phrases such as "faire beau [to be fine (weather)]") that express a function or notion. These

phrases are selected by awarding organizations based on subjective criteria such as perceived usefulness or topic relevance but not corpora-based frequency or collocation data. Similarly, the lists for German contain many compound nouns (e.g., "Fahrradverleih [bicycle hire]"), where two or more nouns join to create a new word. However, the extent to which students should be familiar with the components of these multiword phrases or German compounds is not specified. For instance, it is not clear whether students are expected to learn "ça dépend [it depends]" as a fixed phrase only or "Fahrradverleih [bicycle hire]" as a compound or whether they are also expected to understand and productively manipulate the individual components (e.g., "ça [it]" and "dépendre [to depend]" or "Fahrrad [bike]" and "Verleih [hire]"), where neither are listed as isolated lexical items. This leaves some ambiguity surrounding what should be taught and could be assessed.

Many of the entries on these topic-driven lists, moreover, are unlikely to occur outside the specific contexts in which they are presented (e.g., food items such as "choux de bruxelles [brussels sprouts]," "chou-fleur [cauliflower]," and "concombre [cucumber];" see AQA, 2016, pp. 23−85, and Pearson Edexcel, 2018, pp. 73−149, for more examples). Even though the notion of word frequency is acknowledged, albeit indirectly, by the organizations in sections entitled "general vocabulary" (AQA, 2016, p. 13), or "high-frequency language" (Pearson Edexcel, 2018, p. 24), there are no definitions of frequency or parameters surrounding the proportion of high- and low(er)-frequency language that can be used in the lists or exams. As such, general and high-frequency vocabulary has been selected, to the best of our knowledge, using judgments from the awarding organizations' subject experts and relevant stakeholders in the education community (AQA, 2016; Pearson Edexcel, 2018) and not corpus-informed frequency lists.

The analyses presented in this article serve to inform our understanding of the potential value of frequency-informed words lists, a proposal that was under consultation during the period of analysis (see Department for Education, 2021, for the consultation process; Department for Education, 2022a, for the revised curriculum). This proposal received mixed reactions from the foreign language education community and continues to be widely debated even now, even though many teachers have welcomed the changes (Department for Education, 2022b). The revised curriculum specifies that awarding organizations must publish word lists containing 1,200 entries at foundation and 1,700 entries at higher, of which 85% must be high frequency. These lists must be used when creating exams. The lists can also contain up to 50 additional entries from any frequency band, including 30 multiword phrases of up to 5 words and 20 cultural and geographical entries (Department for Education, 2022a, p. 7). The core set of 1,200 and 1,700 entries must be listed in line with Bauer and Nation's (1993) Level 2 word families (i.e., the headword and its inflections). A small number of the high-frequency entries, however, have to be listed as Level 1 word families (i.e., as individual word forms, such as "suis [am]"), as they are highly idiosyncratic forms that are likely to be stored holistically within the lexicon, at least in the early stages of establishing morphosyntactic systems (Pliatsikas & Marinis, 2013). The definition of word families is also extended to include derived forms of the headwords for the reading test only, with the specific derivational morphology defined in the curriculum.

At the time of writing, awarding organizations have developed word lists to fit these parameters for French, which have been approved by Ofqual. Similar lists for German and Spanish are currently under development. Thus, for the current study, we created frequency-informed lists that adhered to the proposed parameters described above. Having outlined the educational context, we now review wider issues related to word lists for language education.

Use of (frequency-informed) word lists in language curricula

The use of word lists is not unique to foreign language education in England. Many providers of international qualifications and governments have published word lists, or at least specified vocabulary learning targets, for their English proficiency tests and language curricula (e.g., MEXT [Japanese Ministry of Education], 2016; Ministry of Education of the People's Republic China, 2017). Word

TABLE 1 Summary of corpus-based, high-frequency lists in English.

List	Words	Families[a]	(F)lemmas[b]	Affixes	Coverage (%)
BNC/COCA3000 (Nation, 2012)	19,062	3,000	9,132	81	90[c]
Nuclear Family List 7 (Cobb & Laufer, 2021)	7,293	3,000	5,610	22	85[c]
New General Service List (Brezina & Gablasova, 2015)	5,115	N/A	2,494	N/A	89[c]
New General Service List (Browne, 2014)	8,342	N/A	3,000	N/A	86[c]
Essential Word List (Dang & Webb, 2016a)	N/A	N/A	800	N/A	75

[a]The headwords and inflected and main derived forms.
[b]Headwords and inflected forms. [c] As reported in Cobb & Laufer (2021, p 859).

lists have also been used for tests of languages other than English, such as the German proficiency tests administered by the Goethe-Institut (2016) and the Japanese proficiency tests administered by Japan Educational Exchanges and Services (2009). We were not, however, able to ascertain the extent to which these lists are frequency informed and/or have been actively used in test or material development.

Nevertheless, the move toward frequency-informed approaches to word selection has been ongoing for some time within the field of English language teaching and research (see Coxhead, 2011, for more discussion). There are several examples of current practice where corpus-based word frequency data together with teachers' and other stakeholders' subjective judgments have guided word selection. One such example is the Hong Kong Education Bureau's (2021) English language curriculum, which includes frequency-informed lists of up to 3,500 words for 15-year-olds and 5,000 words for 18-year-olds. These word lists were developed in collaboration with the Chinese University of Hong Kong, with close reference to word frequency information from the General Service List (West, 1953), the British National Corpus (Nation, 2006a), and the Academic Word List (Coxhead, 2000) and expert judgments from primary and secondary school teachers. Cambridge English (Lanes et al., 2019) has also for some time combined corpora-informed word frequency data with expert judgments to develop word lists to create assessments and enable students and teachers to focus on the vocabulary needed for exams. Finally, for languages other than English, frequency-informed word lists have been used to create Dutch proficiency exams in the Netherlands (College voor Toetsen en Examens, n.d.) and promoted as part of the French language curriculum in primary schools in France (Éduscol, 2020).

Why is word frequency a useful selection principle? The importance of coverage

The move toward stipulating a core of high-frequency words aligns with findings that knowledge of the first 2,000 most frequently occurring words provides at least 82% coverage of written language and 89% of spoken language in English (Dang & Webb, 2014; Nation, 2006b; Webb & Nation, 2017; Webb & Rodgers, 2009a, 2009b). These findings, in large part, motivated the development of several influential corpus-based lists of high-frequency words in English. Cross-corpus analyses have found that these lists (summarized in Table 1) can cover up to 90% of the words in large general corpora of written and spoken language in North American and British varieties of English.

Given that learners need to know at least 95% of the words in any given written or spoken text for unassisted comprehension (Hu & Nation, 2000; Laufer, 1998; Laufer & Ravenhorst–Kalovski, 2010; Schmitt et al., 2011), these findings suggest that knowledge of high-frequency words plays a critical role in helping learners to understand the meaning of different texts. Indeed, Coxhead (2017) and Coxhead and Boutorwick (2018) have demonstrated the importance of high-frequency vocabulary in helping learners to understand teacher talk, textbooks, and learning materials. As such, a word list drawn up according to the frequency-informed parameters of the new GCSE curriculum has the potential to prepare students as well as, if not better than, the awarding organizations' current lists, which were designed specifically to prepare students for the exams. However, no research to date has examined whether a frequency-informed list would provide adequate coverage of these high-stakes exams and how this coverage compares to the current topic-driven lists.

Implications of lexical coverage for inferencing

Word list coverage has important consequences for test takers' need to draw on higher-level comprehension strategies such as lexical inferencing (i.e., the ability to work out the meaning of specific unknown words using the available clues in the input). It is well observed that second language (L2) readers find lexical inferencing challenging. For instance, it has been reported that at 90% coverage, intermediate learners can only infer 52% (95% CI [46%, 58%]) of unknown words in a written text (Laufer, 2020), and that this ability varies according to factors such as how many words in the text need to be inferred (Sternberg, 1987) and, critically for the educational context of the current study, language proficiency (Hamada, 2014).

In contrast, comparatively fewer studies have explored L2 inferencing skills in listening. The available research, however, suggests that inferencing is even more difficult in listening than reading for several reasons (van Zeeland, 2014). For instance, L2 listeners may not even notice unknown words in the speech stream, or they may have difficulties understanding the clues needed to successfully infer meaning, given that the speech stream is ephemeral. In contrast, inferencing during reading can be helped by the fact that our eyes can revisit written input. Strategy-based instruction, however, has been shown to improve inferencing as well as confidence (i.e., self-efficacy) in and attitudes toward L2 reading (Graham et al., 2020; Macaro & Erler, 2008) and listening (Graham & Macaro, 2008) among beginner-to-low-intermediate classroom learners of French in England. By extrapolation, it could be argued that difficulties with inferencing can have a detrimental impact on learners' confidence in and attitudes toward language learning.

In sum, the level of lexical inferencing required can determine text difficulty for each individual reader or listener. Thus, the extent to which any word list provides coverage of the exams indirectly informs us about the likely need for test takers to infer meaning.

Topic-driven curricula and semantic relatedness in the early stages of learning

Whether topics or frequency inform word selection may, of course, wash back into pedagogy. A heavily topic-structured word list is likely to send a message that curricula and pedagogy must be arranged around topics, with teaching planned in semantic clusters of words and phrases. In contrast, an entirely frequency-informed list may not facilitate strong semantic clustering, at least in the early stages of learning. By definition, topic-specific vocabulary is not usually high frequency, and so there may simply be insufficient topic-specific vocabulary to create clusters when learners' lexicons are small. Therefore, if it were the case that semantic clustering was clearly beneficial for learning at these early stages, it would perhaps be unwise to invoke a frequency-based lexical selection principle.

However, the value of semantic clustering in fact remains opaque. Several laboratory studies report negative effects of semantic clustering. These effects include learners needing more time to acquire

(Tinkham, 1993, 1997; Waring, 1997) and translate (Finkbeiner & Nicol, 2003) first (L1)–L2 word combinations that were grouped according to semantic categories than ones that were not. Similar difficulties have been observed in classroom studies (Erten & Tekin, 2008; Karabulut & Dollar, 2016). The influence of semantic clustering may also in part depend on the extent to which visual (physical) features are shared between the referents of semantically similar words. For instance, Ishii (2015) found that sets of words describing objects with physical similarities were more difficult to learn than unrelated words and semantically related words without physical similarities.

Nevertheless, two classroom studies (Hashemi & Gowdasiaei, 2005; Hoshino, 2010) have reported benefits of learning words in semantically related sets compared to unrelated sets. Two laboratory studies have also suggested some limited benefits of semantic clustering, but these benefits were only short lived or very specific. For instance, Schneider et al. (2002) found initial facilitatory effects when learners were tested from the L2 to the L1, but this effect disappeared only 1 week after the intervention. Kemp and McDonald (2021) further observed that semantic clustering only facilitated learning when tests were from L2 to the L1. In other cases, semantic clustering either had no effect or hindered learning. In sum, the extent to which semantic clustering improves learning is unclear at best and, in many cases, seems not to be beneficial.

The current study cannot address this debate, as it is not an intervention study. We mention it here because it seems valuable to investigate the potential value of a frequency-based lexical selection principle given the lack of clear evidence in favor of semantic clustering. Furthermore, and perhaps more importantly, the lexical selection (i.e., frequency- or topic-driven) principle does not in any case fully determine whether semantic clustering can or cannot happen in curricula and pedagogy. That is, frequency-informed approaches do not preclude some thematic grouping of words. Curricula can still provide communicative contexts, scenarios, or themes, even for beginner-to-low-intermediate learners, and such thematic grouping becomes increasingly easier as more words are known. Indeed, the new GCSE curriculum requires awarding organizations to "identify a limited number of broad themes or topics with relevance to the countries or communities where the language is spoken" (Department for Education, 2022a, p. 5).

THE CURRENT STUDY

The current study set out to explore (a) the extent to which words from the awarding organizations' topic-driven lists appear in high-stakes French, German, and Spanish exams, and (b) the extent to which a new frequency-informed word list might prepare learners for the exams relative to the awarding organizations' lists. For these analyses, we calculated overall lexical coverage in order to ascertain the percentage of the words in the exams covered by the headwords and their family members included on the different word lists (Nation & Waring, 1997). This technique is commonly used in word list evaluation studies (Brezina & Gablasova, 2015; Browne, 2014; Dang & Webb, 2016b; Gilner & Morales, 2010; Nation, 2004), where lists are "evaluated on different corpora from the corpus from which they are made" (Nation, 2016, p. 130). Unlike traditional word list evaluation studies, our study evaluates the different lists against a single corpus (of GCSE exams) in order to explore the implications that word list coverage may have specifically for GCSE learners within the context of the research–policy–practice interface. Our study therefore addresses the following two research questions (RQs):

RQ1. To what extent are the words from awarding organizations' current topic-driven word lists used in the current GCSE listening and reading exams?
  - RQ1a. What percentage of the lists are used in the corpus of exams?
  - RQ1b. What percentage of the lists are used in an average exam?
  - RQ1c. What percentage of the lists are used in every exam?
  - RQ1d. What percentage of the lists are used only once in the corpus of exams?

RQ2. How much coverage of the GCSE listening and reading exams is provided by the awarding organizations' current topic-driven word lists and frequency-informed word lists?
  RQ2a. How much coverage do the lists provide of the corpus of exams?
  RQ2b. How much coverage do the lists provide of words from flemmas used in every exam?
  RQ2c. How much coverage do the lists provide of words from flemmas used only once in the corpus of exams?

METHOD

The corpus of exam texts

We analyzed a corpus containing 116,647 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 of entry (foundation and higher), and two modalities (listening and reading). At the point of analysis, the corpus included all the exams published by the two organizations for the Department for Education's (2015) curriculum, as exams were not produced in any subject in 2020 or 2021 due to the pandemic and the 2022 exams fell outside the analysis period.

Before profiling the exams, we removed all rubrics, instructions, and comprehension questions in English from the texts. The comprehension questions in the target language were included in the corpus because they were part of what is assessed by the current exams.[1] Proper nouns were retained, given their inclusion as entries in dictionaries (e.g., countries and cities), frequency values in many corpora (Kilgarriff et al., 2014), and lack of reliable transparency for learners. Compound nouns in German not listed as entries in the FreeLing 3.0 dictionary (Padró & Stanilovsky, 2012) were split using CharSplit (Tuggener, 2016), an ngram-based compound splitter for German.

Procedure: lexical profiling tool (MultilingProfiler)

The 96 exams were profiled using the MultilingProfiler (Finlayson et al., 2022, http://multilingprofiler.net/) software. For each exam, the language (French, German, or Spanish) was selected from the "Language" drop-down menu, the frequency list from the "List" menu, and top 5,000 words from the "Level" menu. Each exam text was then copied and pasted into the profile window and the "Download Stats (.csv)" button pressed to extract flemma-based lists of words in the text. Like lemmas, flemmas include a headword and its inflected forms, but unlike lemmas, they do not take the part of speech into account (Bauer & Nation, 1993; Webb, 2021). For instance, "sourire [smile]" and "sourires [smiles]" is one noun lemma, and "sourire [to smile]" and its inflections (e.g., "souris," "souri") is one verb lemma; "sourire" as a flemma includes both of these lemmas. The flemma-based lists provided data about each headword and its members, including its number of occurrences within the text profiled 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 the position of each flemma within corpora-based frequency lists of the 5,000 most frequently occurring (f)lemmas in the respective languages (Davies & Davies, 2017, for Spanish; Lonsdale & Le Bras, 2009, for French; Tschirner & Möhring, 2019, for German). Note that since the analyses were completed for the current study, an updated version of the MultilingProfiler has become available. As such, the analyses may not all be fully reproducible).

The word lists

We analyzed two types of word lists: topic-driven and frequency-informed.[2] The topic-driven lists were developed by the awarding organizations as part of their specifications for the Department for

TABLE 2 Total number of flemmas in each word list.

Language	Tier	AQA	Edexcel	Frequency informed
French	Foundation	1,058	1,811	1,057
	Higher	1,322	2,076	1,496
German	Foundation	1,477	1,926	1,114
	Higher	1,688	2,199	1,546
Spanish	Foundation	1,313	1,797	1,035
	Higher	1,550	2,031	1,458
Mean (*SD*)	Foundation	1,283 (211)	1,845 (71)	1,069 (41)
	Higher	1,520 (185)	2,102 (87)	1,500 (44)

Abbreviation: AQA, Assessment and Qualifications Alliance.

Education's (2015) current curriculum. On average across the three languages, the AQA lists contained a total of 1,283 flemmas at foundation tier and 1,520 at higher tier (see Table 2), of which 51% at foundation and 48% at higher were high-frequency. The Edexcel lists were consistently longer, with 1,845 flemmas at foundation and 2,102 at higher, of which 45% at foundation and 43% at higher were high-frequency.

The frequency-informed lists were developed and evaluated in line with the requirements of the new GCSE curriculum. On average across the three languages, these lists included 1,069 flemmas at foundation and 1,500 at higher, of which 83% were high frequency (equivalent to at least 85% of the entries, as per the requirements of the new curriculum). Subjective judgments about word relevance from teachers and materials designers were also involved in the list creation process.

The two list types used different methods to define and categorize words. To take this into account, each list itself was profiled, using the MultilingProfiler in order to standardize (i.e., flemmatize) the lists and thus allow for meaningful comparisons. This approach to standardization also meant that the individual components of multiword phrases were treated, across all lists, as individual flemmas. For instance, "faire beau [to be fine (weather)]" was analyzed as two separate flemmas (i.e., "faire" and "beau") rather than as one combined phrase.

In the revised GCSE curriculum, expectations regarding knowledge of derivational morphology (e.g., affixes) are clearly defined. This is not the case for the current curriculum. Thus, to allow for meaningful comparisons, neither the frequency-informed lists nor the organizations' lists included the derived forms of the flemmas used, even if such forms are permitted by the relevant curriculum or organizations' specifications. For more information about these lists, including how they were developed, see Online Supporting Information B.

Analysis

RQ1 examined the extent to which the word lists have been used in GCSE exams to date. Our aim was to provide a nuanced understanding of how often words are used in ways that are relevant to the different stakeholders in the community, including a student sitting an 'average' exam or using past exams as practice, a teacher preparing cohorts of students for every and all exams, and an awarding organization creating different exams over the years. As such, we calculated the proportion of flemmas from the lists that had been used: (a) at least once across four sets of exams (i.e., in the corpus of exams; RQ1a), (b) in an average exam (RQ1b), (c) in every exam (RQ1c), and (d) only once across four sets of exams (RQ1d). For these analyses, each flemma on the lists was coded as a binary variable according to whether it had been used in an exam or not.[3] Binomial logistic models were then computed for each

sub-RQ (separately for each language and awarding organization), using the in-built glm function in the R environment (R Development Core Team, 2014). Year, tier, and their two-way interaction were included as predictors in the RQ1b models to explore the extent to which word list use in an average exam varied as a function of year or tier. Tier was included as a predictor in the models for RQ1a, RQ1c, and RQ1d to explore the extent to which it moderated the percentage of the word list used. To account for differences in word list length (as we were interested in the lexical selection principles rather than list length), we included the log transformation of the total number of flemmas used in each word list as an offset.

RQ2 compared the coverage that the organizations' topic-driven lists and the frequency-informed lists provided of (a) the corpus of exams (RQ2a), (b) the words from flemmas used in every exam (RQ2b), and (c) the words from flemmas used only once across four sets of exams (RQ2c). Binomial logistic models were computed for each sub-RQ (separately for each language and awarding organization), with the log transformed word list length as an offset. To model coverage of the flemmas and their members in the exams, each observation (i.e., flemma) was weighted by how many times its members appeared within the relevant set of exams. List type, tier, modality, and their three-way interaction were included as predictors in the models for RQ2a (for the corpus of exams) to investigate the extent to which coverage varied as a function of list type, tier, or modality. List type, tier, and their two-way interaction were included as predictors in the models for RQ2b (in every exam) and RQ2c (only in one exam) in order to explore the extent to which the effect of word list type varied as a function of tier. We did not include modality in these models because the RQ2a analyses provided sufficient insight into the effect of modality, and further analyses were beyond the scope of this article.

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 and each contrast comparing one level against the reference level (i.e., list type: the frequency-informed list vs. the awarding organization's list; tier: higher vs. foundation; modality: reading vs. listening). For RQ1a, the actual exams were compared against the sample exam to investigate the extent to which the sample exam was representative of an actual exam sat by students.

Model summaries were calculated, using the tab_model function from the sjPlot package (Lüdecke, 2021), and included (unstandardised) odd ratios (ORs), 95% confidence intervals (CIs), and p values for each predictor, and Nagelkerke's psuedo R2 to assess the fit of the model.

Significance was evaluated at an alpha level of 0.05. Provided in brackets are ORs, including 95% CIs and *p* values, with full model specifications reported in Online Supporting Information C (for RQ1) and D (for RQ2). Where relevant, the emmeans (Lenth, 2021), ggeffects (Lüdecke et al., 2021), and ggplot (Wickham et al., 2021) packages were used to probe any significant interactions. Due to space constraints, descriptive statistics for the Edexcel analyses are presented in Online Supporting Information C and D.[4]

RESULTS

Research question 1: Use of words from the current lists in the exams

Percentage of the lists used in the corpus of exams (RQ1a)

On average across the three languages, 52% of the AQA lists at foundation and 55% at higher were used in the corpus of exams—that is, at least once across four sets of exams (see Table 3). This means that on average, about half of the words on the lists have never appeared in an exam.

The models revealed that these percentages were similar across tiers, languages, and awarding organizations, with one exception in French, where a higher percentage of the AQA lists were used at foundation than at higher, OR: 1.25, 95% CI [1.06, 1.47], $p = .009$.

TABLE 3 Percentage of the Assessment and Qualifications Alliance (AQA) lists used in the corpus of exams.

	Tier	
Language	**Foundation (%)**	**Higher (%)**
French	61	61
German	42	47
Spanish	54	58
Mean (*SD*)	52 (9)	55 (7)

TABLE 4 Mean (*SD*) percentage of the Assessment and Qualifications Alliance (AQA) lists used in an average exam.

	Tier	
Language	**Foundation (%)**	**Higher (%)**
French	34 (1)	34 (2)
German	22 (2)	24 (2)
Spanish	29 (1)	31 (1)
Mean	28 (6)	30 (4)

Percentage of the lists used in an average exam (RQ1b)

On average across the three languages, 28% of the AQA lists appeared in an average exam at foundation and 30% at higher (see Table 4).

Overall, the models revealed that these percentages were similar across the AQA exams, with a few relatively small tier- or year-dependent differences. For instance, in French, a higher percentage of the lists were used at foundation than at higher, OR: 1.30, 95% CI [1.19, 1.41], $p < .001$, and in German, in the actual exams than their sample counterparts, OR: 2018 vs sample: 1.27, 95% CI [1.13, 1.43], $p < .001$; 2019 vs sample: 1.24, 95% CI [1.10, 1.40], $p < .001$; 2020 vs sample: 1.16, 95% CI [1.03, 1.31], $p = .016$.

The mean percentages were lower for Edexcel (see Table C7 in Online Supporting Information), though a slightly different pattern of results emerged in terms of the effects of tier and year. A higher percentage of the lists were used in the German and Spanish (but not French) exams at foundation than higher, OR: German: 1.10, 95% CI [1.03, 1.18], $p = .007$; Spanish: 1.12, 95% CI [1.04, 1.20], $p = .004$. The models also revealed some inconsistencies between the actual and sample exams, observed in four out of nine possible comparisons.

Percentage of the lists used in every exam (RQ1c)

On average across the three languages, 10% of the AQA lists at foundation and 11% at higher were used in each and every exam (see Table 5). The models revealed that these percentages remained constant across tiers for both awarding organizations.

Percentage of the lists used only once across four exams (RQ1d)

On average across the three languages, 21% of the lists at foundation and 23% at higher were only used once in the AQA exams (see Table 6). The models revealed that these percentages were similar across tiers and languages for both awarding organizations, with one exception where a higher percentage of

TABLE 5 Percentage of the Assessment and Qualifications Alliance (AQA) lists used in every exam to date.

	Tier	
Language	**Foundation (%)**	**Higher (%)**
French	13	13
German	8	9
Spanish	10	11
Mean (*SD*)	10 (3)	11 (2)

TABLE 6 Percentage of the Assessment and Qualifications Alliance (AQA) lists used only once across four exams.

	Tier	
Language	**Foundation (%)**	**Higher (%)**
French	22	24
German	19	21
Spanish	22	23
Mean (*SD*)	21 (2)	23 (2)

TABLE 7 Mean (*SD*) coverage of the corpus of Assessment and Qualifications Alliance (AQA) exams.

			List type	
Language	**Tier**	**Number of words**	**AQA (%)**	**Frequency-informed (%)**
French	Foundation	9,069	74	84
	Higher	12,832	76	88
German	Foundation	7,828	68	87
	Higher	10,920	71	88
Spanish	Foundation	7,507	77	85
	Higher	10,983	78	88
Mean (*SD*)	Foundation	8,135 (825)	73 (5)	85 (2)
	Higher	11,578 (1,086)	75 (4)	88 (<1)

the lists were only used once in the German Edexcel exams at foundation than at higher, OR: 1.21, 95% CI [1.04, 1.41], $p = .013$.

Research question 2: Coverage provided by the two list types

Coverage of the corpus of exams (RQ2a)

On average across the three languages, the AQA lists covered 73% of the words in the corpus of exams at foundation and 75% at higher, whereas the frequency-informed lists covered 85% at foundation and 88% at higher (see Table 7).

Once the length of each word list was controlled for, the statistical models revealed that the frequency-informed list—an example of a new GCSE list—consistently provided greater coverage of the GCSE exams than the awarding organizations' current lists, regardless of modal-

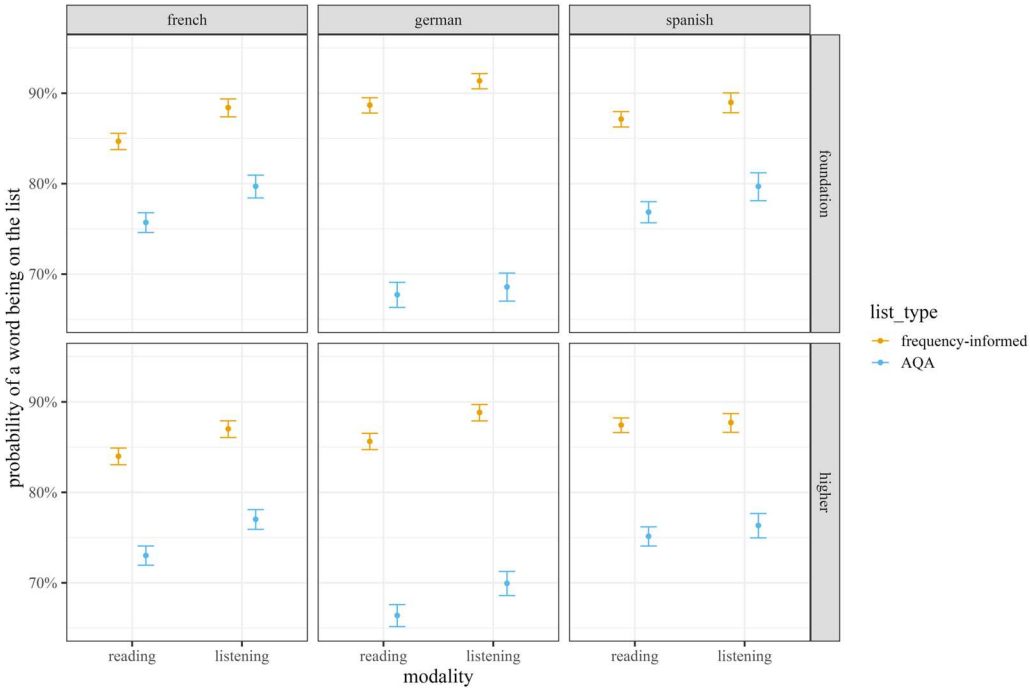

FIGURE 1 Predicted coverage of the corpus of Assessment and Qualifications Alliance (AQA) exams after controlling for word list length. [Color figure can be viewed at wileyonlinelibrary.com]

ity, tier, language, or awarding organization, ORs: AQA French: 1.91, 95% CI [1.82, 2.01], $p < .001$; AQA German: 3.70, 95% CI [3.50, 3.91], $p < .001$; AQA Spanish: 2.15, 95% CI [2.03, 2.28], $p < .001$; Edexcel French: 1.34, 95% CI [1.26, 1.42], $p < .001$; Edexcel German: 1.83, 95% CI [1.72, 1.94], $p < .001$; Edexcel Spanish: 1.76, 95% CI [1.66, 1.88], $p < .001$ (see Figure 1).

In general, controlling for list length, both types of word lists provided greater coverage of the GCSE exams at foundation than at higher, ORs: AQA French: 1.13, 95% CI [1.07, 1.19], $p < .001$; AQA German for the frequency-informed lists only: 1.32, 95% CI [1.21, 1.45], $p < .001$; AQA Spanish for the listening exams only: 1.17, 95% CI [1.07, 1.29], $p = .001$; Edexcel French: 1.28, 95% CI [1.20, 1.36], $p < .001$; Edexcel German: 1.25, 95% CI [1.18, 1.32], $p < .001$; Edexcel Spanish: 1.21, 95% CI [1.14, 1.29], $p < .001$. There were only two exceptions to this pattern. One was where the AQA lists provided similar coverage of both tiers of the German exams. The other was where both types of word lists provided similar coverage of both tiers of the AQA Spanish reading exams.

Another largely consistent finding was the effect of modality. In most cases, once list length was controlled for, both lists provided greater coverage of the listening than the reading exams, ORs: AQA French: 1.29, 95% CI [1.22, 1.36], $p < .001$; AQA German: 1.22, 95% CI [1.15, 1.29], $p < .001$; AQA Spanish at foundation only: 1.19, 95% CI [1.09, 1.30], $p < .001$; Edexcel French: 1.10, 95% CI [1.04, 1.17], $p = .002$; Edexcel Spanish: 1.09, 95% CI [1.03, 1.16], $p = .006$. There were only three exceptions: two (i.e., in the AQA Spanish and Edexcel German exams, both at higher) where both lists provided comparable coverage of the listening and reading exams, and one (i.e., in the Edexcel German exams at foundation) where both lists provided greater coverage of the reading than the listening exams, OR: 1.15, 95% CI [1.05, 1.25], $p = .002$.

TABLE 8 Mean (*SD*) coverage of words from flemmas used in every Assessment and Qualifications Alliance (AQA) exam.

Language	Tier	Mean (*SD*) number of words averaged across years	List type	
			AQA (%)	Frequency-informed (%)
French	Foundation	1,599 (243)	85 (1)	98 (<1)
	Higher	2,334 (295)	87 (1)	98 (<1)
German	Foundation	1,387 (132)	75 (1)	99 (<1)
	Higher	1,947 (216)	80 (1)	99 (<1)
Spanish	Foundation	1,247 (138)	88 (1)	99 (<1)
	Higher	1,923 (135)	88 (2)	99 (<1)
Mean	Foundation	1,411 (221)	83 (6)	98 (1)
	Higher	2,068 (283)	85 (4)	99 (1)

TABLE 9 Coverage of words from flemmas used only once across four sets of Assessment and Qualifications Alliance (AQA) exams.

Language	Tier	Mean (*SD*) number of words averaged across years	List type	
			AQA (%)	Frequency-informed (%)
French	Foundation	260 (12)	28	25
	Higher	339 (50)	30	36
German	Foundation	216 (27)	37	32
	Higher	310 (31)	34	33
Spanish	Foundation	235 (19)	38	30
	Higher	292 (29)	37	37
Mean (*SD*)	Foundation	237 (26)	34 (6)	29 (4)
	Higher	314 (39)	33 (4)	35 (2)

Coverage of words from flemmas used in every exam (RQ2b)

On average across the three languages, the AQA lists covered 83% of the words from flemmas used in each and every exam at foundation and 85% at higher, whereas the frequency-informed lists covered 98% of these words at foundation and 99% at higher (see Table 8).

The models confirmed that the frequency-informed lists consistently provided greater coverage of these words than the AQA lists, OR: French: 6.63, 95% CI [5.91, 7.46], $p < .001$; German: 30.98, 95% CI [26.31, 36.77], $p < .001$; Spanish: 13.97, 95% CI [11.75, 16.74], $p < .001$, regardless of language or tier. A similar pattern emerged for the Edexcel exams and their lists.

Coverage of words from flemmas used only once in the corpus of exams (RQ2c)

On average across the three languages, the AQA lists covered 34% of the words from flemmas used only once in the corpus of exams at foundation and 33% at higher, and the frequency-informed lists covered 29% of these words at foundation and 35% at higher (see Table 9). Post-hoc comparisons revealed no differences in the coverage between the lists, regardless of tier.

For Edexcel, some similar results were found, albeit with a few potential differences. For German and Spanish, the two lists provided similar coverage of words used only once at higher, but at foundation, the Edexcel lists provided slightly greater coverage than the frequency-informed lists, OR: German: 1.34, 95% CI [1.11, 1.62], $p = .002$; Spanish: 1.29, 95% CI [1.05, 1.57], $p = .014$. For French, the Edexcel lists generally provided greater coverage of the words used only once than the frequency-informed lists. However, the difference was only slight with a lower-bound 95% CI of 1, OR: 1.14, 95% CI [1.00, 1.30], $p = .047$.

DISCUSSION

We now discuss the results in light of the broader aims of the study.

To what extent is teaching from the awarding organizations' word lists an effective use of curriculum time in terms of preparation for the GCSEs?

The awarding organizations' word lists were created as guides to plan lessons and write textbooks in prepartion for the GCSE exams. To date, however, we have not known the extent to which teaching from these lists is an effective use of limited curriculum time. Our analyses (in response to RQ1) revealed that in an average exam, AQA have used approximately 28% of their lists at foundation and 30% at higher. Perhaps more sobering are findings that in four sets of exams, only 52% of the lists at foundation and 55% at higher have been used, only 10% of the lists at foundation and 11% at higher have been used in every exam, and over one fifth of the lists (21% at foundation and 23% at higher) have been used only once across four sets of exams. This means that approximately 48% of the lists at foundation and 45% at higher could be consuming learning effort during a 2-year GCSE course, but not empowering students with the relevant knowledge and skills for these high-stakes exams. This time could perhaps be better spent elsewhere, for example, by revisiting words used more frequently in the exams so that learners can retrieve these words more easily during exams and in other contexts of use.

To what extent do the awarding organizations' topic-driven lists prepare students for the GCSE exams?

On average across the three languages, the AQA lists covered 73% of the words in the corpus of exams at foundation and 75% at higher. Once we controlled for the length of each list, our analyses (in response to RQ2) largely showed that the current lists provided greater coverage of the foundation than the higher exams and of the listening than the reading exams, regardless of awarding organization. These findings are to some extent expected. First, differences in coverage between tiers might reflect differences in proficiency levels, with word lists providing more support for lower proficiencies, on the assumption that the higher exams are more likely to require greater lexical knowledge and/or inferencing skills because they contain a higher proportion of words off the list. Second, listening is typically considered to be more challenging than reading. For instance, in listening, learners are less able to identify known words (van Zeeland, 2013) and infer the meaning of unknown words (van Zeeland, 2014). As such, it is probably reasonable for listening texts to contain a higher proportion of words from the lists (i.e., more likely to be known).

This coverage is some way below the coverage provided by other (f)lemma-based word lists, such as Browne's (2014) and Brezina and Gablasova's (2015) New General Service Lists (89% and 86%, respectively). Both of those lists, however, are considerably longer (at 3,000 lemmas) than the awarding organizations' lists, and so the lower coverage is perhaps to be expected. However, this same

argument cannot fully explain the lower coverage, because Dang and Webb's (2016a) Essential Word List only contained 800 lemmas and yet provided 75% coverage (see Table 1). Critically, this coverage is substantially below the 95% to 98% of words that learners need to know in any given text to fully understand it (Hu & Nation, 2000; Laufer, 1998; Laufer & Ravenhorst–Kalovski, 2010; Schmitt et al., 2011).

This insufficient coverage could be a consequence of organizations not actively using their lists when creating the exams or the lists not containing a high enough proportion of high-frequency words, given that high-frequency words would normally represent over 80% of any written or spoken text (Dang & Webb, 2014; Nation, 2006b; Webb & Nation, 2017; Webb & Rodgers, 2009a, 2009b). Corroborating the latter explanation are Dudley and Marsden's (2023) findings that 44% of the flemmas used in the corpus of exams at foundation and 45% at higher are indeed low(er)-frequency (defined in that study as beyond the 2,000 most frequent words). Relatedly, it may be that the lower coverage observed in our study is in part due to the high proportion (49% at foundation and 52% at higher) of low(er)-frequency flemmas on the AQA lists (see Online Supporting Information B for more information). As such, awarding organizations may find it challenging to include more of these low(er)-frequency items on their lists in any single exam, especially given (a) the short length of the texts in the exams, (b) the inevitably high proportion of high-frequency words needed for any given text, and (c) the breadth of topics covered by the lists.

Taken together, these findings indicate that the lists currently used to inform teaching and textbooks do not reliably equip learners with the relevant lexical knowledge for the exams. The limited use of the word lists in assessments, compounded by their optionality, suggests that the lexical content of these high-stakes exams is not "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). Indeed, as noted earlier, Pearson Edexcel (2020) acknowledged that "spoken extracts [and texts] contain too many words not on vocabulary lists" (p. 11). Thus, learners are perhaps unlikely to have been exposed to these "off list" words a sufficient number of times to be able to notice and understand them in exams, given that some studies suggest that a learner needs to be exposed to a word at least 10 times before they can retrieve its meaning with ease (Brown et al., 2008; Waring & Takaki, 2003). These findings also broadly align with Webb and Paribakht's (2015) and Jin et al.'s (2016) findings relating to a mismatch between what is taught and what is assessed in terms of the lexical content of high-stakes exams.

In sum, our analyses suggest that a significant proportion (up to 25%) of the lexical items used in these exams do not come from the word lists and thus cannot be predicted. The lack of predictability in turn incurs a greater emphasis on lexical inferencing skills, which can vary greatly among learners, contexts, and modalities (Hamada, 2014; Laufer, 2020). Thus, the unpredictability of the lexical content may be contributing to the perceived and actual difficulty of these high-stakes exams.

Does a frequency-informed word list provide similar or better coverage of GCSE exams relative to the current topic-driven lists?

Our analyses (in response to RQ2) showed that the frequency-informed word lists, created according to the parameters of the new GCSE curriculum, provided without exception greater coverage of the words used in the corpus of exams (85% at foundation and 88% at higher) than the (AQA) lists (73% at foundation and 75% at higher), across tiers, modalities, and languages. Similarly consistent findings were observed for coverage of words from flemmas used in every exam. Critically, this stronger coverage pertained even though the exams that we analysed were not created in line with the frequency-informed lists—such exams would only be created for a future GCSE exam.

The only exception to this pattern of results was among the very small set of words from flemmas used only once across four sets of exams where the awarding organizations' lists provided similar coverage, or slightly higher coverage in a small number of cases, relative to the frequency-informed

lists. These findings are not surprising given that these rarely used words are, by definition, more likely to be low(er) frequency, unpredictable, and topic specific.

More generally, our findings could perhaps be accounted for by an argument that the frequency-informed lists simply contain a greater proportion of the extremely high-frequency words that are used in every exam but that the organizations intentionally chose to omit from their lists on the (unstated, covert) assumption that students have prior knowledge of these words from earlier in their education. We emphasize, however, that this is not an adequate explanation because the organizations' lists do, in fact, already cover a very high proportion (83% at foundation and 85% at higher for AQA) of this relatively small set of high-frequency words used in every exam. Instead, our findings are likely to be better explained by the fact that the organizations underuse their word lists when creating exams, as presented earlier, and discussed later.

Admittedly, however, neither the topic-driven nor the frequency-informed lists provided sufficient coverage for unassisted comprehension of the current exams, which is thought to require between 95% and 98% coverage (Schmitt et al., 2011; van Zeeland & Schmitt, 2013). Nevertheless, a frequency-informed list would provide sufficient coverage for future GCSE exams as those assessments would be created using frequency-informed lists similar to the lists developed for the current study. We believe that such a step could increase the chances that comprehension depends less on guessing the meaning of unfamiliar words and more on knowledge of a (more) realistic amount and type of language that is appropriate to a GCSE. However, further research using the awarding organizations' word lists and exams developed under the revised curriculum is needed to verify this.

We are not suggesting that learning words from a frequency-informed list is any easier than learning words from a topic-driven list. Indeed, high-frequency words are often difficult to learn due to factors such as their "semantic neutrality, length, part of speech, polysemy, morphological [ir]regularity, cognateness, [and] orthographic transparency" (Hashimoto, 2021, p. 182). Furthermore, most learners are unlikely to learn all 1,250 (at foundation) or 1,750 (at higher) items on the new lists, given that receptive vocabulary knowledge at this stage has been estimated to be between 550 and 850 words (David, 2008; Milton, 2006, 2015). Therefore, even with a prescribed word list, inferencing skills would still be assessed as individuals attempt to understand unfamiliar words.

LIMITATIONS AND FUTURE DIRECTIONS

The current study is not without limitations. As explained in the method section, the word lists analyzed in this study were flemma-based. This means that (a) a very small number of words were assigned to the incorrect family, and (b) any derived forms of the flemmas were not included even if such derived forms were permitted by the relevant curriculum. It is therefore likely that the coverage reported in this study may have (slightly) underestimated true coverage for both lists. Future analyses could compare the differences in coverage when the permitted derived forms are considered.

Throughout this article and in our previous work, we have suggested that the extensive use of low(er)-frequency words and the limited use of awarding organizations' word lists in the GCSE exams may, in part, be contributing to the perceived and actual difficulty of languages at GCSE and, ultimately, to declines in the number of students studying languages. Although our data provide partial, indirect support for these possibilities, further research with GCSE learners is needed to investigate the extent to which actual knowledge of the lexical content of the GCSE exams—including test takers' understanding of the comprehension questions themselves—predicts (a) performance in listening, reading, writing, and speaking, and (b) the decision to continue studying languages at A level and beyond.

A reviewer raised the issue of whether or not the underuse of the lists has been deliberate. We speculate not for two reasons. First, it has not been easy for awarding organizations to check whether the exams align with their word lists—at least until now—due to the unavailability of lexical profiling software in German and Spanish (see, however, Cobb's Lextutor [http://www.lextutor.ca] for French). Second, the organizations have not been required to use their word lists when creating the exams, or

even document which words they have used. It may even be that they felt compelled to underuse their own lists to create exams that were not too long while also meeting the requirement to include words off their lists in exams (Ofqual, 2021a). This latter point, however, is unlikely to be a bona fide or comprehensive reason, as simply providing shorter lists and/or lists with a greater proportion of high-frequency words would have reduced the underuse. Of course, without undertaking further research (e.g., by interviewing the test developers), there is no way of knowing whether the underuse has been unintentional or perhaps intentional to elicit a normal distribution of assessment data.

CONCLUSION

Taken together, we argue that the awarding organizations' current word lists are not optimally serving students or teachers for several reasons. First, the lists do not provide sufficient coverage of the listening and reading exams to allow for adequate comprehension. As such, the exams contain a substantial amount of what is likely to be unfamiliar language. Second, valuable curriculum time is likely spent teaching words that have never appeared in the exams, which seems out of kilter with how assessments are expected to cover the entire curriculum over 3−5 years (Ofqual, 2021b). We suggest that these factors may have, in part, contributed to the perceived and actual difficulty of the exams and, over time, detrimentally impacted the number of students studying languages at GCSE and beyond (University Council of Modern Languages, 2021).

Our study suggests that frequency-informed lists might help to mitigate these difficulties to some extent, as they consistently provided greater coverage of the listening and reading exams relative to current lists. We found that frequency-informed lists may, in fact, better prepare learners for exams in a topic-driven curriculum relative to the awarding organizations' own topic-driven lists. However, this does not imply that lexical selection should be driven solely by frequency, as word usefulness or topic relevance is also important (Dang et al., 2020), as reflected in the creation of the frequency-informed lists used here. What may wash back into curricula and pedagogy from a frequency-informed word list is a reduction in the number of very context-specific words that learners are unlikely to need in assessment or life. Such words are perhaps better learned on an as-needed basis through personal experiences, including study, travel, or work.

We note that although changes to the GCSE curriculum may help tighten the link between what is taught and what is assessed, any influence this may have on the achievement–motivation cycle is a topic for future research. Nevertheless, we hope that these findings have begun to address Dang et al.'s (2020) observation that alongside word list evaluation studies (Cobb & Laufer, 2021), research is needed to better understand the potential use of such lists in the context of curriculum design and/or high-stakes assessment.

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. We would like to thank colleagues at the former National Centre for Excellence for Language Pedagogy (now Language-Driven Pedagogy) for preparing the exam papers and word lists for profiling and creating illustrative examples of a frequency-informed GCSE word list; 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). We are also grateful to the three anonymous reviewers, Prof. Marta Antón as *MLJ* Editor, and Dr. Kate Miller for their insightful feedback, comments, and editing work on the manuscript. All remaining errors are our own.

CRediT author statement: **Prof. Emma Marsden**: conceptualization (lead); methodology (supporting); formal analysis (supporting); investigation (supporting); resources (lead); writing – original draft (supporting); writing – review & editing (lead); supervision (lead); project administration (lead); funding acquisition (lead). **Dr. Amber Dudley**: conceptualization (supporting); methodology (lead); formal analysis (lead); investigation (lead); resources (supporting); data curation (lead); writing – original draft (lead); writing – review & editing (lead); visualization (lead); project administration (lead). **Dr. Rachel Hawkes**: conceptualization (supporting); resources (supporting); writing – review & editing (supporting); funding acquisition (supporting).

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/hvfja/?.

ORCID

Emma Marsden https://orcid.org/0000-0003-4086-5765
Amber Dudley https://orcid.org/0000-0003-2904-9150

ENDNOTES

[1] A reviewer quite rightly pointed out that the vocabulary in the comprehension questions may determine comprehension difficulty. Although of clear importance for future research, such an analysis was beyond the aims of the current article for several reasons. To accurately determine the source of difficulty (i.e., question vs. text), it would be critical to check learners' actual comprehension of these parts separately, rather than rely on coverage data alone. Furthermore, in future exams, all GCSE comprehension questions will be in English. As such, analyzing the vocabulary in the comprehension questions did not align with the current aims of determining the usefulness of different list types.

[2] Available on our OSF repository: https://osf.io/hvfja/?

[3] See R Scripts on our OSF repository: https://osf.io/hvfja/?

[4] Full datasets and analyses, including R scripts, are provided on our OSF repository (https://osf.io/hvfja/?)

REFERENCES

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

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

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

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*, 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*, 1–22. https://doi.org/10.1093/applin/amt018

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*, 136–163. http://hdl.handle.net/10125/66816

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

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., & Laufer, B. (2021). The nuclear family word list: A list of the most frequent family members, including base and affixed words. *Language Learning*, *71*, 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*, 462–480. https://doi.org/10.1093/applin/amw019

College voor Toetsen en Examens. (n.d.). *Wat is het Staatsexamen Nt2?* [*What is the NT2 state exam?*]. https://www.staatsexamensnt2.nl/over-het-staatsexamen-nt2/wat-is-het-staatsexamen-nt2

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

Coxhead, A. (2011). The academic word list ten years on: Research and teaching implications. *TESOL Quarterly*, *45*, 355–362. https://doi.org/10.5054/tq.2011.254528

Coxhead, A. (2017). Academic vocabulary in teacher talk: Challenges and opportunities for pedagogy. *Oslo Studies in Language*, *9*, 29–44. https://doi.org/10.5617/osla.5845

Coxhead, A., & Boutorwick, T. J. (2018). Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science. *TESOL Quarterly*, *52*, 588–610. https://doi.org/10.1002/tesq.450

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*, 66–76. https://doi.org/10.1016/j.esp.2013.08.001

Dang, T. N. Y., & Webb, S. (2016a). Making an essential word list for beginners. In I. S. P. Nation (Ed.), *Making and using word lists for language learning and teaching* (pp. 153–167). John Benjamins. https://doi.org/10.1075/z.208.15ch15

Dang, T. N. Y., & Webb, S. (2016b). Evaluating lists of high-frequency words. *ITL—International Journal of Applied Linguistics*, *167*, 132–158. https://doi.org/10.1075/itl.167.2.02dan

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

David, A. (2008). Vocabulary breadth in French L2 learners. *Language Learning Journal*, *36*, 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. (2021). *GCSE MFL subject content review*. https://consult.education.gov.uk/ebacc-and-arts-and-humanities-team/gcse-mfl-subject-content-review/

Department for Education. (2022a). *French, German and Spanish: GCSE subject content (January 2022)*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1046043/GCSE_French_German_Spanish_subject_content.pdf

Department for Education. (2022b). *French, German and Spanish GCSE subject content review: Government consultation response*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1046071/FINAL_government_consultation_response_-_SP_Jan_2022.pdf

Dudley, A., & Marsden, E. (2023). *The number and frequency of words 16-year-olds need for their French, German, and Spanish exams*. OSF Preprints. https://doi.org/10.31219/osf.io/pzqkm

Eduqas. (2019). *Eduqas GCSE French specification*. https://www.eduqas.co.uk/media/0gqg4xeh/eduqas-gcse-french-spec-from-2016-e.pdf

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

Ellis, R. (2019). Towards a modular language curriculum for using tasks. *Language Teaching Research*, *23*, 454–475. https://doi.org/10.1177/1362168818765315

Erler, L., & Macaro, E. (2011). Decoding ability in French as a foreign language and language learning motivation. *Modern Language Journal*, *95*, 496–518. https://doi.org/10.1111/j.1540-4781.2011.01238.x

Erten, I. H., & Tekin, M. (2008). Effects on vocabulary acquisition of presenting new words in semantic sets versus semantically unrelated sets. *System*, *36*, 407–422. https://doi.org/10.1016/j.system.2008.02.005

Finkbeiner, M., & Nicol, J. (2003). Semantic category effects in second language word learning. *Applied Psycholinguistics*, *24*, 369–383. https://doi.org/10.1017/S0142716403000195

Finlayson, N., Marsden, E., & Anthony, L. (2022). MultilingProfiler (Version 2) [Computer software]. University of York. https://www.multilingprofiler.net/

Garon–Carrier, G., Boivin, M., Guay, F., Kovas, Y., Dionne, G., Lemelin, J.-P., Séguin, J. R., Vitaro, F., & Tremblay, R. E. (2016). Intrinsic motivation and achievement in mathematics in elementary school: A longitudinal investigation of their association. *Child Development*, *87*, 165–175. https://doi.org/10.1111/cdev.12458

Gilner, L., & Morales, F. (2010). Corpus-based frequency profiling: Migration to a word list based on the British National Corpus. *The Buckingham Journal of Language and Linguistics*, *1*, 41–57. https://doi.org/10.5750/bjll.v1i0.3

Goethe-Institut. (2016). *Goethe-Zertifikat A2: Wortliste [Goethe Certificate A2: Wordlist]*. Goethe-Institut. https://www.goethe.de/pro/relaunch/prf/en/Goethe-Zertifikat_A2_Wortliste.pdf

Graham, S., Courtney, L., Tonkyn, A., & Marinis, T. (2016). Motivational trajectories for early language learning across the primary-secondary school transition. *British Educational Research Journal*, *42*, 682–702. https://doi.org/10.1002/berj.3230

Graham, S., & Macaro, E. (2008). Strategy instruction in listening for lower-intermediate learners of French. *Language Learning*, *58*, 747–783. https://doi.org/10.1111/j.1467-9922.2008.00478.x

Graham, S., Woore, R., Porter, A., Courtney, L., & Savory, C. (2020). Navigating the challenges of L2 reading: Self-efficacy, self-regulatory reading strategies, and learner profiles. *Modern Language Journal*, *104*, 693–714. https://doi.org/10.1111/modl.12670

Hamada, M. (2014). The role of morphological and contextual information in L2 lexical inference. *Modern Language Journal*, *98*, 992–1005. https://doi.org/10.1111/modl.12151

Hashemi, M. R., & Gowdasiaei, F. (2005). An attribute-treatment interaction study: Lexical-set versus semantically-unrelated vocabulary instruction. *RELC Journal*, *36*, 341–361. https://doi.org/10.1177/0033688205060054

Hashimoto, B. J. (2021). Is frequency enough?: The frequency model in vocabulary size testing. *Language Assessment Quarterly*, *18*, 171–187. https://doi.org/10.1080/15434303.2020.1860058
Hawkes, R., & Lillington, C. (2016a). *Viva! AQA GCSE (9-1) Spanish foundation student book*. Pearson Education.
Hawkes, R., & Lillington, C. (2016b). *Viva! AQA GCSE (9-1) Spanish higher student book*. Pearson Education.
Hawkes, R., & Lillington, C. (2016c). *Viva! Edexcel GCSE (9-1) Spanish foundation student book*. Pearson Education.
Hawkes, R., & Lillington, C. (2016d). *Viva! Edexcel GCSE (9-1) Spanish higher student book*. Pearson Education.
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
Hong Kong Education Bureau. (2021). *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
Hoshino, Y. (2010). The categorical facilitation effects on L2 vocabulary learning in a classroom setting. *RELC Journal*, *41*, 301–312. https://doi.org/10.1177/0033688210380558
Hu, H.-C., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403–430.
Ishii, T. (2015). Semantic connection or visual connection: Investigating the true source of confusion. *Language Teaching Research*, *19*, 712–722. https://doi.org/10.1177/1362168814559799
Japan Educational Exchanges and Services. (2009). *New Japanese-language proficiency test guidebook: Executive summary*. https://www.jlpt.jp/e/reference/pdf/guidebook_s_e.pdf
Jin, T., Li, Y., & Li, B. (2016). Vocabulary coverage of reading tests: Gaps between teaching and testing. *TESOL Quarterly*, *50*, 955–964. https://doi.org/10.1002/tesq.324
Karabulut, A., & Dollar, Y. K. (2016). The effects of presenting different types of vocabulary clusters on very young learners' foreign language learning. *Education 3–13*, *44*, 255–268. https://doi.org/10.1080/03004279.2014.904391
Kemp, L. S., & McDonald, J. L. (2021). Second language vocabulary acquisition: The effects of semantic relatedness, form similarity, and translation direction. *Language Learning*, *71*, 716–756. https://doi.org/10.1111/lang.12449
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*, 7–36. https://doi.org/10.1007/s40607-014-0009-9
Lanes, A., Love, R., Kalman, B., Brenchley, M., & Pickles, M. (2019). Updating the A2 Key and B1 Preliminary vocabulary lists. *Cambridge Assessment English - Research Notes*. https://www.cambridgeenglish.org/Images/561337-key-preliminary-revisions-wordlists.pdf
Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *19*, 255–271. https://doi.org/10.1093/applin/19.2.255
Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*, *54*, 1076–1085. https://doi.org/10.1002/tesq.3004
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*, 15–30.
Lenth, R. (2021). *emmeans: Estimated marginal means, aka leastsquares means.*. R package version 1.7.1.1. https://cran.r-project.org/web/packages/emmeans/index.html
Lonsdale, D., & le Bras, Y. (2009). *A frequency dictionary of French*. Routledge. https://doi.org/10.4324/9780203883044
Lüdecke, D. (2021). *sjPlot: Data visualization for statistics in social science, R package version 2.8.10. R Package Version 2.8.10*. https://cran.r-project.org/web/packages/sjPlot/index.html
Lüdecke, D., Aust, F., Crawley, S., & Ben–Shachar, M. S. (2021). *ggeffects: Create tidy data frames of marginal effects for "ggplot" from model outputs (1.1.1)*. https://strengejacke.github.io/ggeffects/
Macaro, E., & Erler, L. (2008). Raising the achievement of young-beginner readers of French through strategy instruction. *Applied Linguistics*, *29*, 90–119. https://doi.org/10.1093/applin/amm023
MEXT. (2016). *Eigo kyouiku ni kan suru heisei 29 nendo gaisan youkyu to nitsuite* [*Concerning Heisei 29 requirement for English education*]. http://www.mext.go.jp/b_menu/shingi/chukyo/%0Achukyo3/004/siryo/_icsFiles/afieldfile/2016/10/28/1378911_3.pdf
Milton, J. (2006). Language lite? Learning French vocabulary in school. *Journal of French Language Studies*, *16*, 187–205. https://doi.org/10.1017/S0959269506002420
Milton, J. (2015). French lexis and formal exams in the British foreign language classroom. *Revue Francaise De Linguistique Appliquee*, *20*, 107–119. https://doi.org/10.3917/rfla.201.0107
Ministry of Education of the People's Republic China. (2017). *English curriculum standards for senior high schools*. People's Education Press.
Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Newbury House.
Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/cbo9781139524759
Nation, I. S. P. (2004). Vocabulary learning and intensive reading. *English Australia Journal*, *21*, 20–29. https://doi.org/10.26686/wgtn.12560303.v1
Nation, I. S. P. (2006a). *The BNC word family lists*. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs

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

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

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

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy*. (pp. 6–19). Cambridge University Press.

Ofqual. (2019). *Inter-subject comparability in GCSE modern foreign languages*. https://www.gov.uk/government/news/inter-subject-comparability-in-gcse-modern-foreign-languages

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

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. In N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, *Proceedings of the Eighth Language Resources and Evaluation Conference* (pp. 2473–2479). European Language Resources Association.

Parrish, A., & Lanvers, U. (2019). Student motivation, school policy choices and modern language study in England. *Language Learning Journal*, *47*, 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. (2020). *French, German & Spanish: A guide to the amendments to our assessments from summer 2021*. https://qualifications.pearson.com/content/dam/pdf/GCSE/Modern-Languages/MFL_GCSE_Assessment_Amendments_Guide.pdf

Pliatsikas, C., & Marinis, T. (2013). Processing of regular and irregular past tense morphology in highly proficient second language learners of English: A self-paced reading study. *Applied Psycholinguistics*, *34*, 943–970. https://doi.org/10.1017/S0142716412000082

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

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

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

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*, 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*, 484–503. https://doi.org/10.1017/S0261444812000018

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*, 419–44. https://doi.org/10.1006/jmla.2001.2813

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.

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. *Modern Language Journal*, *98*, 902–920. https://doi.org/10.1111/modl.12146

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

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System*, *21*, 371–380. https://doi.org/10.1016/0346-251X(93)90027-E

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research*, *13*, 138–163. https://doi.org/10.1191/026765897672376469

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

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

University Council of Modern Languages. (2021). *Report on granular trends in modern languages in UCAS admissions data, 2012–18*. https://university-council-modern-languages.org/wp-content/uploads/2021/07/UCML-BA-UCAS-Granularity-Report.pdf

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

van Zeeland, H. (2014). Lexical inferencing in first and second language listening. *Modern Language Journal*, *98*, 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*, 457–479. https://doi.org/10.1093/applin/ams074

Vu, T., Magis-Weinberg, L., Jansen, B. R. J., van Atteveldt, N., Janssen, T. W. P., Lee, N. C., van der Maas, H. L. J., Raijmakers, M. E. J., Sachisthal, M. S. M., & Meeter, M. (2022). Motivation-achievement cycling in learning: Al literature review and research agenda. *Educational Psychology Review*, *34*, 39–71. https://doi.org/10.1007/s10648-021-09616-7

Waring, R. (1997). The negative effects of learning words in semantic sets: A replication. *System*, *25*, 261–274. https://doi.org/10.1016/S0346-251X(97)00013-4

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*, 130–163.

Webb, S. (2021). The lemma dilemma. *Studies in Second Language Acquisition*, *43*, 941–949. https://doi.org/10.1017/S0272263121000784

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

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test? *English for Specific Purposes*, *38*, 34–43. https://doi.org/10.1016/j.esp.2014.11.001

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

Webb, S., & Rodgers, M. P. H. (2009b). The lexical coverage of movies. *Applied Linguistics*, *30*, 407–427. https://doi.org/10.1093/applin/amp010

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

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D., & RStudio. (2021). *ggplot2: Create elegant data visualisations using the grammar of graphics (3.3.5)*. https://cran.r-project.org/web/packages/ggplot2/index.html

Wilkins, D. A. (1976). *Notional syllabuses*. Oxford University 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: Marsden, E., Dudley, A., & Hawkes, R. (2023). Use of word lists in a high-stakes, low-exposure context: Topic-driven or frequency-informed. *Modern Language Journal*, *107*, 1–24. https://doi.org/10.1111/modl.12866