TO CITE THIS CHAPTER:

Marsden, E., & Hawkes, R. (2023). Situating practice in a limited-exposure, foreign languages school curriculum. In Y. Suzuki (Eds.) *Practice and Automatization in Second Language Research: Theory, Methods, and Pedagogical Implications.* Chapter 4. Routledge.

Abstract for the publisher's website

This chapter describes the design of a language-driven curriculum and how it was informed by some of the theoretical and empirical research on practice. This large-scale, state-funded project aimed to situate effective practice within an engaging curriculum for 11-16-year-olds in England with approximately 400-450 hours of instruction over five years in French, German, and Spanish. In Part 1, we describe the foreign language context and outline the curriculum and pedagogy design tasks we undertook. In Part 2, we describe the extent to which we embedded principles of practice into class materials and professional development. We highlight the affordances that research offered our decision-making and acknowledge some challenges faced in working at the interface between research, policy, and practice. In our Concluding Remarks, we expose and discuss in detail some areas in which we found our research knowledge-base to be severely lacking for informing real-world problems of this nature.

Chapter 4. Situating practice in a limited-exposure, foreign languages school curriculum

Emma Marsden[1] and Rachel Hawkes[2]

[1] Department of Education, University of York, York, UK, [2] The Cam Academy Trust, Cambridgeshire, UK

Introduction

There has been a boom in research into practice, evidenced by this book, a special issue (Suzuki, Nakata, & DeKeyser, 2019), a large meta-analysis (Kim & Webb, 2022), and a growing set of partial and conceptual replications (e.g., Suzuki, 2017); these are all signs of a maturing agenda. In this chapter, we attempt to describe how some of this

burgeoning research fed into addressing a 'real world problem' (Brumfit, 1995) that we and our colleagues were faced with in 2019: to situate effective practice within a curriculum for 11-16-year-olds who have approximately 400-450 hours of classroom instruction spread over five years of mainstream state-school education, in three foreign languages (French, German, and Spanish). This challenge demanded multi-faceted and high-stakes decisions, which were driven by a complex mix of practical constraints and relevant research. After three and a half years, the team[1] at The National Centre for Excellence for Language Pedagogy (NCELP)[2] have made freely available 1,296 fully resourced lessons either for 'off the shelf' use or teachers' to adapt, for teaching 11-14-year-olds. We have essentially produced the equivalent of about nine 'free textbooks', one for each of three languages for three years of lessons (see resources.ncelp.org)—but all of these materials are fully adaptable, for non-commercial purposes, as all are held under a CC BY-NC-SA 4.0 licence. This means that teachers can directly edit them to suit their purposes. These class and homework resources are accompanied by dozens of professional development resources, many of which draw on research and refer to open accessible summaries held on OASIS (Marsden et al., 2018; https://oasis-database.org).

In Part 1 of this chapter, we describe the foreign language context in which we are working and outline the curriculum and pedagogy design tasks we undertook. In Part 2, we describe our approach to embedding principles of practice into resources and professional development, and highlight the affordances that research into effective practice offered our thinking but also acknowledge some of the challenges we faced working at the interface of research, policy, and practice. In our Concluding Remarks, we expose and discuss in detail some areas in which we found our research knowledge-base to be severely lacking for informing real-world problems.

---

[1] The work presented in this chapter draws on that of the NCELP team of resource developers, teacher educators, and researchers: Inge Alferink, Ivan Avaca, Nicholas Avery, Louise Bibbey, Giulia Bovolenta, Louise Caruso, Amber Dudley, Natalie Finlayson, Susan Graham, Victoria Hobson, Amanda Izquierdo, Rowena Kasprowicz, Heike Krüsemann, Ciarán Morris, Catherine Morris, Charlotte Moss, Stephen Owen, Jack Peacock, Hilary Potter, Catherine Salkeld, Kirsten Somerville, Mary Richardson, Janine Turner, Peter Watson, Robert Woore.

[2] NCELP is funded by a grant from Department for Education in England (Dec. 2018- Dec. 2022). It has also received funds from the University of York's *Economic and Social Research Council's Impact Acceleration Fund*, the *British Academy* (for OASIS-database.org), *Research England* (in part funding the database resources.ncelp.org), and the University of York's *Research Development Fund*.

Part I: Context

Background to the Real-World Challenge: Embedding a practice-based pedagogy and curriculum

The National Centre for Excellence for Language Pedagogy (NCELP; ncelp.org) was commissioned by the Department for Education for England to increase the confidence and capacity of teachers to deliver a set of pedagogical recommendations laid out in the Modern Foreign Languages Pedagogy Review (Teaching Schools Council, 2016; non-peer-reviewed, so-called 'grey literature'). The recommendations were essentially underpinned by a broadly defined 'skill acquisition' (information processing) account of learning (DeKeyser, 2017), within cognitive-interactionist perspectives on second language learning (Collins & Marsden, 2016) and classroom learning more generally (Horvath, Lodge, & Hattie, 2016). The report aimed to (re-)focus teachers' attention on the need to plan and sequence language content to give students ample time to establish, consolidate, and automatize their knowledge of core components of language competence: sounds-spelling relations, the lexicon, and morphosyntax. Such a theoretical framework (rather than families of learning theories that foreground roles for implicit learning mechanisms—whether they are domain-general or language-specific theories) seems likely to be an appropriate account of learning in the context at hand, given that: the amount of exposure to the language is low; the content to be covered explicitly defines the knowledge (grammar, vocabulary, and, in 2022, sound-spelling relations) and the skills (listening, reading, writing, speaking) to be learned (DfE, 2015, 2022); and students learn intentionally in a formal school environment where other subjects are taught, learned, and assessed explicitly and intentionally. Critically, neither the Review nor NCELP's work has dismissed a role for implicit learning and knowledge or incidental learning. However, the work discussed in this chapter lays out one approach to the early stages of a mass-education program by drawing on research into practice schedules (i.e., detailed plans of when specific language—lexis, morphosyntax, sound-writing systems—is practised and revisited in classwork and homework) which are designed in order to promote *intentional* learning.

Another major aim of NCELP is to increase the number of students choosing to study a foreign language, due to high rates of drop out from the subject at ages 14, 16, and 18. That is, there is an aspiration that improving curriculum and pedagogy might increase students' sense of—broadly defined—autonomy, intrinsic motivation, self-efficacy, and desire to engage with the culture and people who speak the language by tapping into the achievement-motivation cycle (e.g., Mills, Pajares, & Herron, 2007).

In trying to address these two overarching aims, the parameters that NCELP could *not* change included: an average of just 400-450 hours instruction (with negligible and unreliable hours of instruction in primary school[3]); minimal exposure to language outside school; and the majority monolingual anglophone national context sitting against the global background of English as the major lingua franca.

The challenge was substantial. Developing a curriculum that both motivates and provides sufficient practice for near beginner and pre-intermediate learners, with a very wide range of individual differences (both cognitive, such as analytic ability, and affective, such as motivation) would be no mean feat.

The learners. The diversity of learners was one characteristic that marked this project apart from many research studies into practice effects, the vast majority of which have been conducted in adult populations of university-based students (though see Kasprowicz & Marsden, 2018; Kasprowicz, Marsden & Sephton, 2019; Marsden, 2006; Marsden & Chen 2011). In this context, almost the whole population of pupils in each year, approximately 500,000 children aged 11-13—studies at least one (sometimes two) Modern Foreign Languages (MFLs) as an obligatory part of their education. Approximately 50% of the whole school population (250,000) students then continue to study a MFL aged 14-16 and then take a high stakes, national examination (the General Certificate of Secondary Education [GCSE]), at the end of their secondary school course aged 16. Given the size of this population, and given the fact that school education is organised according to year groups (i.e., age; not proficiency or previous language experience), there is huge diversity in the full range of individual differences between learners (e.g., knowledge, proficiency, motivations, language learning aptitudes, working memory capacity, first language skills, and, for an average of about 20% of the population, their home and/or community languages). The curricula and pedagogy we developed had to offer *illustrative* models, including practice schedules in each language, to teachers who need to cater to this exceptionally heterogenous population of learners.

Broad characteristics of the curriculum and instruction to date. Prior to NCELP and in schools not currently drawing heavily on NCELP's resources, paper textbooks, which are non-obligatory and available commercially, are used. These textbooks, and the schools' own schemes of work (i.e., the detailed teaching plans that

[3] A maximum potential is approximately 80 hours at primary school but most often, secondary school teachers 'start again' with the MFL curriculum.

sequence *what* is to be taught and *when*) derived from the textbooks, are arranged in topics (for examples, Harrison, et al., 2016; Hawkes & Lillington, 2016; see also analyses of textbooks by Häcker, 2008 and Tschichold, 2012). In these, 'practice' has very often been conceptualized as rehearsing situation-based phrases, operationalized in terms of presenting and practising slot and filler frames. These frames are selected on the grounds that they are useful for comprehension and production of the language needed for specific topics, functions and notions (e.g., telling the time, giving an account of a past or planned holiday, arranging to go out). Grammar is usually presented in textbooks in decontextualized sections presenting single or multiple paradigms (agreement, tense, syntax) and practiced in exercises such as written gap-fill type drill style (akin to DeKeyser's 2010 'narrow' conceptualization of practice). In some contexts, pattern spotting (inductive) learning can also be promoted. Somewhat separately, conversations and written production about specific topics (e.g., preferences about food, travel, or leisure time) are often rote-learnt. Listening and reading texts are used to practice holistic comprehension.

Critically, however, the lexicon, grammar, and sound-writing systems are not revisited in planned cycles *across* this topicalized curriculum, so schemes of work do not cumulatively document which or when vocabulary, grammar, or sound-spelling relations have been practiced and revisited. As one best-selling textbook publisher said to me "we do not know how many times we have been using high frequency or low frequency words in our textbooks" (personal communication). Instead, any revisiting of vocabulary and grammar would largely be dependent on revisiting the rote-learned formulae (which of course host vocabulary and grammar) that, in turn, is often achieved by revisiting a specific topic or rehearsed conversation (e.g., revisiting the topic of holidays or free time to practice the past tense; talking about your daily routine to practice reflexives). As Mitchell wrote in 2002 (still relevant twenty years on) this has led to "prioritising and rewarding the rote learning of fixed phrases and the accumulation of vocabulary for use in slot and filler patterns" (p. 19).

In sum, re-visiting of the *language* content itself has largely been reliant upon opportunities arising from a (personal and social) topic-driven curriculum and the associated texts, as opposed to the approach we adopted at NCELP by designing a language-driven curriculum *from which* a wide range of themes (or contexts of use) emerge. From download statistics from the resources portal (currently about 270,000), attendance at and feedback from professional development events, and the size of our mailing lists, we estimate that hundreds (possibly a couple of thousand) of teachers now use and adapt the language-driven NCELP schemes of work and accompanying resources; that is, very detailed practice schedules that lay out precisely when language

is revisited and its functions/contexts of use, and are meticulously linked—via embedded urls—to the relevant full classroom and homework resources for each week of lessons. To our understanding, these teachers are adapting these resources to their own contexts and, in doing so, are moving away from the more topic-driven materials. Although precise numbers and outcomes are currently unknown, we are greatly encouraged by an increasing number of anecdotal reports from teachers (see NCELP Testimonials: https://ncelp.org/testimonials/) who have voluntarily adopted the schemes of work and report that their pupils can create their own meaning and appear to do so more readily than in previous cohorts.

Part II: Principles of Effective Practice

In Part II, we discuss how our curriculum design and resources reflect five principles of L2 practice: (1) deliberate, (2) systematic, (3) challenging, (4) transfer-appropriate, and (5) feedback. For each, we refer to theoretical and empirical research that informed—bottom-up—the curriculum and resources. We also refer to research that is reflected in the resources; that is, teacher preferences or practical constraints largely drove decisions that were nevertheless broadly compatible with the principles.

Deliberate Practice (and Identifying the Content)

Arguably, for practice to be deliberate, the content (the *what* is deliberate) needs to be identified *a priori*, and so the type of the syllabus is key. An analytic syllabus identifies tasks or thematic content (for task-based, content-based, or immersion style pedagogies) and the 'analysis' of language is largely left to the learner and to reactive, focus-on-form instruction. This can sometimes be accompanied by the identification of learning or communication strategies to help learners when their needs outstrip their knowledge. In such contexts, practice means practicing a 'task' (writing an article, giving a talk, debating, making an arrangement) or demonstrating substantive knowledge (about an author, genre, understanding and discussion of a theme or issue). In contrast, a synthetic syllabus defines and sequences the *language* content, which might take substantive content *as a vehicle* (e.g., a Spanish festival) and/or be supported by defined tasks (responding to a newspaper article), but the accumulation of language knowledge would not be *driven* by that substantive or task-based content.

Creating synthetic syllabi seemed appropriate for our purposes for many reasons. First, synthetic syllabi are able to provide a convenient map for practice-driven instruction, as the content can be intuitively divided into language components that can be revisited, spaced, or interleaved (though see Kim et al., 2020 and Suzuki, 2021 that

examined the practice of holistic tasks, but among university students). Second, a synthetic syllabus is more in line with existing practice in schools (even though the content is synthesised by topic, with grammar and vocabulary hooked on). Third, there has been relatively little research on analytic syllabi, such as task-based programmes, in contexts similar to ours. Without a solid evidence base, the risks of applying such a syllabus in a high-stakes and large-scale government investment would have been considerable. Fourth, there is strong consensus that language knowledge components that are identified in synthetic syllabi, such as vocabulary, grammar, phonology, are powerful predictors of proficiency and skills (e.g., for listening, Wallace, 2022). Finally, a synthetic syllabus *can* accommodate functions/notions and be supported by tasks (see Lambert, this volume, for a discussion of Task Supported Language Teaching), so those pedagogical options were not closed.

In view of all this, our practice schedules, that sit within our schemes of work, were primarily driven by descriptions of the language content to be deliberately practiced: the declarative knowledge (i.e., 'knowledge about') of word meaning, sound-symbol correspondences, and grammatical systems that express function and meaning — the knowledge that precedes proceduralized knowledge (i.e., new knowledge that allows a learner to do something with [actually use] the declarative knowledge), or automatized knowledge (i.e., knowledge that is quickly and reliably accessible, without conscious effort; see Chapter 1 for more discussion of the different types of knowledge conceptualised in Skill Acquisition Theory). To determine this language content, decisions had to be made about the lexical, grammatical, phonological and orthographic subsystems that might be feasibly learnt by and useful for learners in our context. We now describe how the declarative knowledge was identified, as this is a critical step for any kind of practice, yet one that has been somewhat neglected in the literature on practice to date.

Defining the lexical content for deliberate practice. Decisions had to be made about how many words one might expect learners to know after 400-450 hours of instruction over five years (based on two hours per week in the first three years, and two and a half to three hours in the last two years, for about 38 weeks per year). What evidence was available to help decision making? Average receptive vocabulary knowledge for learners of French at this level has been estimated at between 550 and 850 word families, with a maximum of 1,800 words known (David, 2008; Milton, 2006, 2015; Milton & Meara, 1998). In terms of the number of words needed for the General

Certificate of Secondary Education[4], an average total of 1,753 different word families is used to create *four* listening and reading examination papers at Higher tier and 1,351 at Foundation tier (Dudley & Marsden, under revision). Other evidence includes a small number of estimates about how many words can be learned in a given amount of time: a general estimate of 10 words per hour's lesson (Schmitt, 2000, p. 144); 3.8 to 4.3 words for every contact hour when learning French in schools in England (Milton & Meara, 1998, p. 75); an average of about 4.5 words learnt per lesson (based on Milton's 2006 findings); and Milton (2015) suggested that 1,500 words for GCSE, usually taken after about 400-450 hours, might be appropriate.

Clearly, the evidence base was thin and opaque; yet, in order to develop example practice schedules for teachers we had to make estimations about how many new words should be introduced each week. We moved ahead on the basis of about 10 lexical items (mainly single words but very occasionally short, multiword phrases) per week (i.e., five items per one-hour lesson). This set is revisited three weeks after it is first introduced and again approximately nine weeks after it was first introduced (i.e., six weeks after its second revisit), making an approximate 0-3-9 week core schedule for the lexicon. Following this, all words are then revisited about six or seven more times over the following two years, mostly intentionally (with explicit practice built into the schedule) but sometimes incidentally (with words being used in activities but not a focus of deliberate practice). This means that after three years (years 7, 8, and 9, known as 'Key Stage 3'), students would have the opportunity to intentionally practice about 950 words (families of inflected but not derived forms). We envisaged that this spacing and revisiting would give time in the curriculum to: practice both productive and receptive knowledge; incorporate intentional learning of polysemous words; learn highly idiosyncratic irregular word forms as separate lexical (holistic, undecomposed) items (see Gor, 2010); learn some multi-word units (Uchihara et al., 2021); and practice some mid-low frequency words given their cultural, personal, social importance.

The practice schedules for the following two years (to complete the five-year school course) were complicated by the existence of two tiers—Foundation and Higher— of final assessment. Under the recently introduced new GCSE qualification (DfE, 2022), Foundation tier exams can include words from a list of 1,250 lexical items (with a reduced set of grammar relative to the Higher tier). Higher tier exams must be created from a list of 1,750 lexical items. For both of these lists, 30 items can be multiword phrases and 20 items must refer to cultural or geographical terms. In

[4] The GCSE is the high stakes external national examination taken by students aged 16, prior to selecting three Advanced (A) levels which in turn lead to University entrance.

addition, reading tests may include true and close cognates and words created from a defined set of derivational morphology. Thus, for our own curriculum design, once these parameters were taken into account, *about* 300 / 800 additional words were left to be learnt for Foundation and Higher tier students respectively over the final two years of the five-year course, as well as consolidating and building depth of knowledge of the 950 words taught previously.

To date, the choice of 'which' words has been determined by assumptions about what might be useful for learners of this age, informed by notions of functional/notional usefulness, among teachers, textbook writers, and the awarding organisations who create the exams. In our selection, we (additionally) drew on objective data about usefulness, i.e., corpora-informed frequency information. We also needed to consider different parts of speech and, in particular, the importance of a verb lexicon, important for driving morphosyntactic development (with tentative evidence, from this learning context, provided by Marsden & David, 2008). As noted above, our process of word list creation was closely informed by the development of the new GCSE Subject Content (DfE, 2022) that was happening in parallel with NCELP's evolving work. This determined that 85% of the word list (to be precise, 85% of 1,700 of the lexical items at Higher Tier, and 1,200 of those at Foundation tier) should be selected from the top 2,000 most frequent word families; the remaining items were selected according to perceived usefulness and relevance, as rated by educators and materials designers (for more details, see Marsden, Dudley, & Hawkes, under revision; Finlayson, Marsden, & Hawkes, in progress). However, an obvious challenge we faced was that as practice schedules progress, as the months and years pass, the word lists become large and unwieldy. So, ensuring that teaching activities (e.g., reading and listening texts) align with both this frequency information and the schedule for introducing words could not be done by hand. Unfortunately, lexical profiling tools in languages other than English are severely lacking (with, to the best of our knowledge, one exception, for French only, Cobb's Compleat Vocabulary Profiler). Thus, we created a freely available lexical profiler (Finlayson, Marsden, Anthony, Bovolenta, & Hawkes, 2021; see https://www.multilingprofiler.net/) that drew on corpus-based frequency information in the three languages. Using this tool, the resource developers could ensure that each set of classroom resources and assessments deliberately drew on planned balances of 'familiar' and 'new' lexical items.

Defining the grammatical content for deliberate practice. The planned 'spine' around which the practice schedules revolve consists of the grammatical subsystems. These grammatical spines in each language are documented in meticulous detail to

enable tracking of the evolving schedule of grammar practice, laying out when each morphosyntactic subsystem is introduced and re-visited. Subparts of inflectional paradigms are revisited in new contexts along several different dimensions: new lexical contexts (i.e., ringing the changes in the lexicon hosting the grammar) as lexical variability is likely to support learning of grammatical regularities (Brooks et al., 2006; Gómez, 2002); juxtaposed against different grammatical subsystems (e.g., present tense singular agreements might be contrasted against past singular agreements; then against present plural agreements, Kasprowicz & Marsden, 2018; Marsden, 2006; Marsden & Chen, 2011; McManus & Marsden, 2019a, 2019b); new semantic contexts (e.g., a city in Canada, or the life of a singer); in different kinds of activities (from controlled to freer); in different modes and modalities (comprehension/production; oral/written), to support transfer appropriate processing (see below).

Grammatical content is presented in the schedule itself by brief descriptions of broad functions that the grammar could serve. For example, in the very initial lessons, parts of copula BE (e.g., first and second person singular) are introduced along with a few regular, high frequency adjectives so that learners can immediately manipulate and practice sentences. S-V inversion for interrogatives in French are introduced along with a small set of new verbs (that follow an already- introduced regular pattern) in the context of asking and answering about what another person does.

For decisions about sequencing—the order of grammar features—we could not really draw on research, despite decades of research into developmental sequences (see Hulstijn, 2015, for a range of different perspectives), the nature of grammatical difficulty (DeKeyser, 2016) and complexity (Housen, et al., 2019). From this literature, we extracted broad principles about the nature of difficulty and complexity, with a view to informing teachers' expectations about learners' progress and which kinds of features may be slowest to develop (if ever), most prone to backsliding. Critically, however, for designing a practice-based scheme of work, it was not clear to us whether it was better for a scheme of work to start with the *most difficult* (as, according to several learning theories – skill acquisition, usage-based, emergentist – such features may need the most classroom attention, practice opportunities, or exposure) or to start with the *least difficult* (as they may be the most motivational and/or be essential precursors for later knowledge, e.g., Keßler, Lenzing, & Liebner, 2016). As defining and corroborating the validity of learnability sequences for all grammatical features in three languages was not within our scope, we could not base our recommendations on the teachability/learnability sequences. Moreover, subparts within a single grammatical subsystem can vary in difficulty – some forms or functions of the grammatical system may meet several of the criteria for 'easy', whereas other parts of the system may be

complex (e.g., have cross-linguistic complexities, multiple meanings in the L2, opaque sound-spelling relations, or low salience). Thus, to inform decisions about sequencing we relied on the team's classroom experience and our broadly research-informed principles: deliberate revisiting of morphosyntax in different lexical and semantic contexts, with expectations about accuracy and fluency tempered according to task characteristics such as time, communicative pressure, familiarity, and cognitive burden.

Defining sound-writing systems content for deliberate practice. To choose which sound-symbol correspondences were appropriate and useful to promote core literacy, we drew on the little existing research into phonics instruction in this context (e.g., Woore, 2011; Woore et al., 2020). That research adopted a 'bang for your buck' principle whereby frequency of the sound-symbol correspondence and its crosslinguistic difficulty informed selection. Similarly, our schemes of work prioritise the most frequently-occurring correspondences, including where they are cross-linguistically complex, whilst rarer (and more complex) correspondences are taught later. A range of approaches to practising phonics is available (Adesope et al., 2011). We opted for an approach whereby a small amount of time every lesson (5-10 minutes) is dedicated to practicing specific sound-symbol correspondences and, later, sounds of the language such as stress patterns. This systematic and planned approach is supplemented by awareness-raising (deliberately incorporated into the lesson materials) to promote noticing of previously learnt correspondences during reading and listening activities. Wherever possible, the lexical and grammatical foci of the lessons overlap with the sound-symbol correspondences taught in that (set of) lessons (for example, stress versus no stress on final o / ó in Spanish conveys person and tense information).

One benefit of our choice to provide synthetic and isolated phonics practice (rather than frontloading phonics at the start of the course, or adopting an entirely analytic, reactive, focus-on-form approach) was that a systematic practice schedule could control how often the phonics content was intentionally practiced. Another advantage was that the high frequency words taught for phonics could reinforce the practice schedule that was developed for the lexicon. However, one downside of isolating phonics in each lesson is that it is not *always* possible to ensure harmony with the grammatical, functional, or semantic foci of each lesson that drive the substantive cohesion in each lesson in order to promote the use of interconnecting sentences about, for example, 'describing when things happen', 'a city in Canada', 'a festival', or 'buying food'. Given that we were unable to locate robust relevant evidence about whether integrated or isolated phonics practice is the most effective, our decision to adopt an isolated (with some integration) approach was driven by the benefits of being

able to produce a practice schedule. A more analytic approach would have run the risk of certain correspondences or phonological features being under-represented over time. In short, research to date could not provide sufficiently robust evidence to inform our choices of 'which' phonics to teach, how to sequence them, or how much time was needed for different kinds of learners. In these respects, practitioners' experience strongly influenced our decisions.

In summary, the 'deliberateness' principle is reflected in many elements of curriculum planning, pedagogy, and assessment (see https://resources.ncelp.org/schemes-of-work). Students engage in deliberate (i.e., conscious, intentional) learning of specific language forms, following synthetic syllabi which identify the lexical, grammatical, and sound-symbol relations for specific lessons. These schemes of work are offered freely as examples of how language knowledge and practice can be sequenced and re-visited systematically to support progression in the early stages of language development within a low exposure foreign language setting. Teachers are encouraged to use or adapt them to inform their own schemes.

Systematic

We interpreted 'systematic' to mean a work plan that has an explicit system driving it. Envisioning and actually creating this system has been perhaps the most intellectually and practically demanding puzzle for the team: how to design a curriculum that broadly adheres to principles of practice by systematically weaving together different components of knowledge whilst ensuring the communication of meaning is kept foremost and space is made for new language in a (broadly defined) expanding practice schedule. Whilst there is some good evidence that expanding practice schedules are beneficial (Gerbier, Toppino, & Koenig, 2015; Nakata, 2015), we are also aware of some (arguably more tentative) evidence that other kinds of spacing can be equally or more beneficial, at least in tests that were administered in the longer-term (e.g., Schuetze, 2015; Schueteze & Weimer-Stuckmann, 2011; see Kim & Webb, 2022 for a review and meta-analysis). Furthermore, in a much more practical sense, during our curriculum development work we observed that there seemed in fact, to be no choice – the schedules had to be 'expanding', at least at higher levels of planning (inter-weeks, i.e., between clusters of lessons) in order to both accommodate new content and also revisit previous language. At a more micro level (within two or three sessions within one week, or within one single session), it seemed undesirable to systematically control whether practice has equal, expanding, or contracting spacing. As Schuetze (2015) noted, "Ideally, a uniform schedule is 100 percent uniform, for example a two–two–two interval. However, the class schedule did not permit for that." (p. 34). It was impractical and unrealistic to control spacing to a fine

degree of minutes within every lesson. For example, within one lesson, sticking to either a schedule of vocabulary practice at 0 then 10 then 20 minutes, or one of 0 then 5 then 20 minutes – would have been very unnatural and demotivating if applied consistently. Also, it would have unduly restricted the accommodation of other content and have led to incoherent lessons that did not facilitate varied activities. Thus, we use the term 'expanding practice schedule' loosely to refer to our system of adding more vocabulary at approximately regularly expanding intervals and revisiting language that had been taught months or years ago. By definition, this meant that the spaces between content (vocabulary, grammar, a particular skill) get larger, as the weeks, months, and years progress.

It has taken teams of experts (including linguists, educators, illustrators, native speakers, and materials editors) nearly three years to create three years (approximately 228 hours) of such materials, in each language, which interleave ever growing bodies of lexical, grammatical, phonological and orthographic (sub)systems in lessons with semantic (thematic) coherence. Interleaving (e.g., ABAB rather than AABB content) was imposed (quasi-)systematically at various levels, from sound-symbol relations through to syntax, given its evidenced benefits for accuracy (Nakata & Suzuki, 2019; though, for effects on fluency, see Suzuki, 2021).

Here we sketch in broad terms the approach we took for vocabulary practice. Our account is necessarily practical given that we could not locate existing research to inform decisions about *precisely* how best to space and when to repeat all of the required content (e.g., words, grammar, sound-spelling relations) for this particular learning context (or any context, for that matter). Instead, these decisions were driven by practical constraints: the total number of lexical items to be learned (1,750), an average (externally determined) total number of lessons (approximately 450 hours), an average fixed maximum time period (5 years), and the rest of the content to be covered (knowledge, skills, embedded in material about the target culture, people and place) which is to a large extent pre-determined by the government's Department for Education (e.g., DfE, 2022).

In week 1, students are given a set of intentional vocabulary learning activities, which can be done using Computer Assisted Language Learning (e.g., Quizlet). The following week, these words are practiced in class, via listening, reading, writing, and speaking activities. These words are revisited about three weeks (or six lessons) later: the students are asked to re-learn them in intentional learning activities in their pre-lesson vocabulary learning homework and many of the words are incorporated into the materials for that lesson. Although not all words can always be incorporated 100% systematically for *intentional* learning in the class activities, they are all *used* in the materials, thus

available for implicit tallying [5] (Ellis, 2002) or consolidation with some level of awareness. In the same way, the vocabulary set is encountered a third time about 5-6 weeks (or 10-12 lessons) later (about nine weeks after the first intentional learning activity). Note that as new language gets added to the schedule, the second and the third intentional learning of a lexical set alongside a brand-new set make a total of about thirty words to be practiced each week. Creating such a schedule requires meticulous record-keeping of when grammar and words are encountered. Figure 1 shows the blue print of the core vocabulary revisiting schedule, at 0, 3, and 9 weeks, as well as the introduction of new grammar.

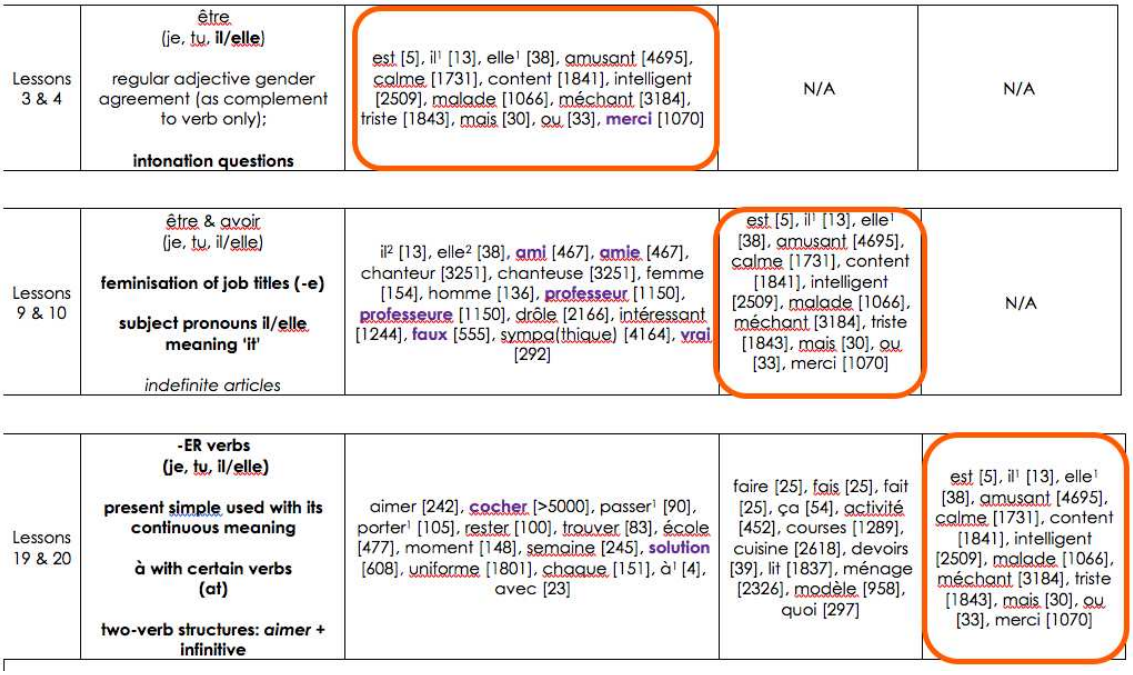

Figure 1: Extract from the French scheme of work for year 7 learners aged 11-12.

Note. The lesson number, with two lessons per week, is shown in the first column. The target grammar is shown in the second column, with bold indicating new grammar. The highlighted boxes indicate the recycled vocabulary. The numbers in square brackets indicate word frequency ranking. The bold words in column 3 (in purple in the actual resource) indicate useful classroom language. The superscript numbers indicate that a word is deemed polysemous for L1 English speakers, with [1] indicating the first meaning introduced (e.g., 'he') and [2] indicating the second meaning (e.g., 'it').

After this initial three-part core schedule, the words are incorporated as

[5]Whereby initial registration via explicit noticing can then be followed by implicit consolidation during subsequent encounters with that form.

intentional learning, for homework and in lessons, in 'consolidation weeks' in the following years. As shown in Figure 2, there are six such consolidation weeks for French in year 8. In addition, previously taught words can also recur as incidental items in lesson resources, and they are explicitly assessed in tests (described below, and see Finlayson & Marsden, 2022). Figure 2 further shows how the same set of words reappears in lessons that are focused on different functional or thematic content. This reflects evidence that variability of semantic context helps learning (e.g., Kemp & MacDonald, 2021), by varying the contexts in which the words are practiced each time.

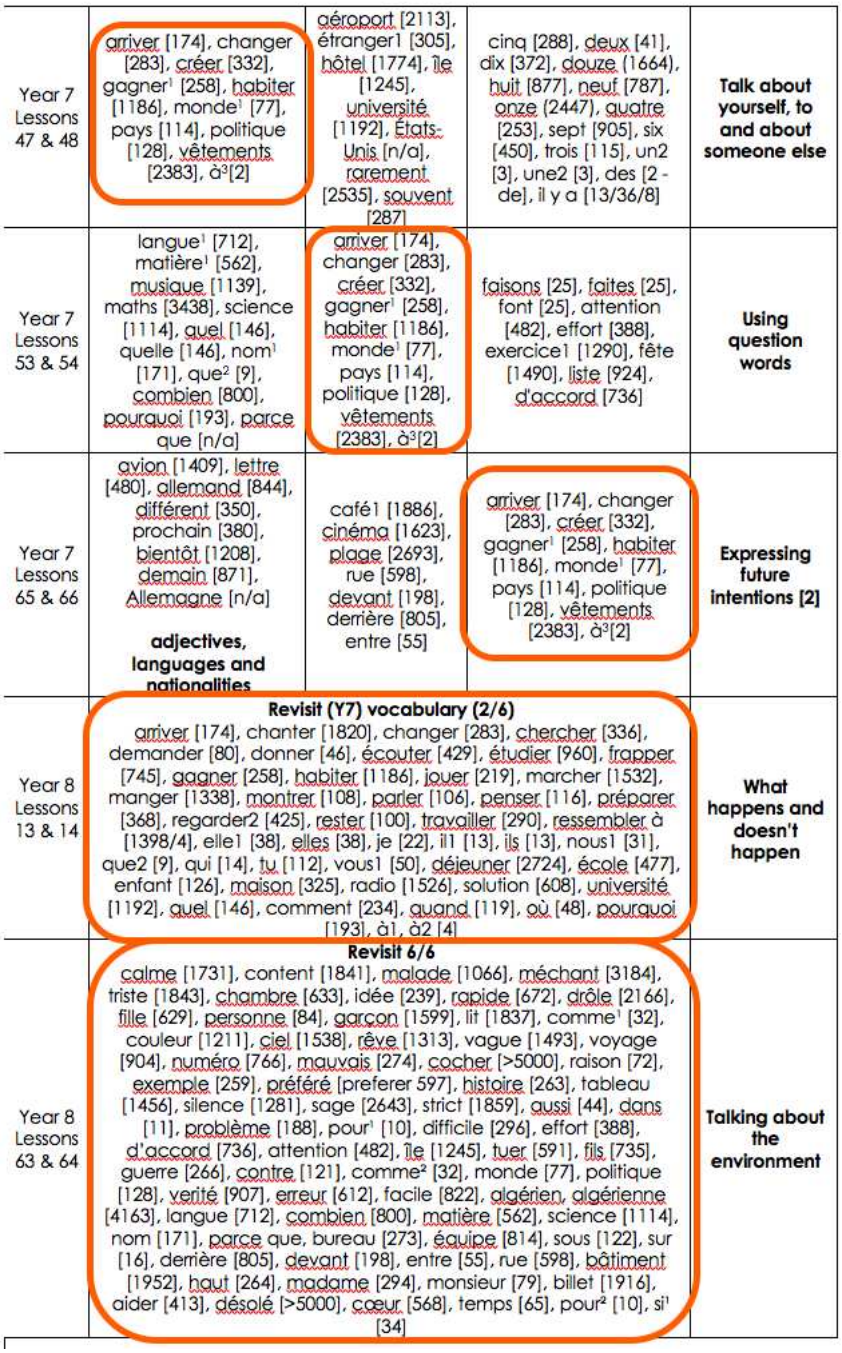

Figure 2: Sets of words in different semantic and functional contexts, from the French scheme of work for years 7 and 8 learners aged 11-13. *Note.* The highlighted boxes show how vocabulary is recycled in the 0-3-9-week schedule and how the vocabulary set then reappears, split over larger sets, in further consolidation practice and in preparation for assessment ('Revisit 2/6' and 'Revisit 6/6' refer to these larger sets being the 2nd and 6th out of 6 'consolidation weeks').

In addition, word families gradually get larger as new grammatical features are introduced, as shown in Figure 3. So, more types (exemplars) of a particular pattern (*-er* verbs in French) are added into the lexicon, as the inflectional paradigm that has been practiced enlarges. Verb paradigms are staggered over several lessons or weeks, to allow interleaving of specific morphemes (e.g., *-s* with *-ent*; *-ons,* with *-ez*).

Lessons 23 & 24	-ER verbs (je, tu, il/elle, **nous**) present simple used with its continuous meaning intonation questions	marcher [1532], manger [1338], préparer [368], **regarder**[1] [425], travailler [290], nous[1] [31], déjeuner [2724], film [848], maison [325], **partenaire** [1077], télé [2746], dehors [1217], préféré [préférer 597]
Lessons 25 & 26	-ER verbs (je, tu, il/elle, nous, **ils/elles**)	chanter [1820], **étudier** [960], jouer [219], ils [13], elles [38], élève [1068], fruit [896], histoire[1] [263], radio [1526], ensemble [124]
Lessons 27 & 28	-ER verbs (je, tu, il/elle, nous, **vous**, ils/elles)	**fermer** [757], **regarder**[2] [425], vous[1] [50], chemise [3892], **classe** [778], **fenêtre** [1604], **porte** [696], **salle** [812], **silence** [1281], **tableau** [1456], **bien** [47]

Figure 3: Practice schedule in the scheme of work showing lexical and inflectional growth. *Note*. Bold in the second column indicates new inflections. Bold in the third column indicates language that is useful for the classroom (purple in our actual resources online). Superscript numbers, on words deemed to be polysemous for L1 English speakers, refer to the 1st or 2nd etc. meaning that has been introduced.

Time is allotted in the schedule for summative testing twice a year. In each round of tests, all words that have been introduced since the last round of tests are assessed, and a sample of 'old' words is tested, reflective of the notion of cumulative testing (Nakata, Tada, McLean, & Kim, 2021). For the first few rounds of the achievement tests, about 50% of the items assessed content that had been taught since the last tests ('recent' material), and the remaining 50% sampled from content that *preceded* the last tests ('old' material). As more content is taught and the body of knowledge-to-be-tested grows, a tension builds between testing old and recent language, given the finite amount of class time available for tests (about one or two hours for the summative tests required by schools). At each test, it becomes necessary to sample fewer items from old (previously

tested) words in order to be able to test all of the *most recent* content. Or, if it were possible to increase the amount of class time given to testing, then the ratio of recent to old items within each test would inevitably tip in favor of old items, to allow the amount of sampling from old items to stay constant as the body of previously taught words gets bigger. But such a situation is unlikely in most classrooms.

The lexical profiler facilitates compatibility checks between tests (and any texts at all) and each week of the schedule. That is, users can select a specific week of the schedule (e.g., Spanish, year 8, term 1, week 5), and the profiler highlights in orange (i.e., indicting a word is 'off the list', or unfamiliar to date) those words that have not yet been encountered and, moreover, any *inflectional morphological* features that have not yet been encountered. That is, words introduced by a specific point in the schedule can carry the inflections that have been encountered up to that same point. (The profiler also allows users to compare their texts against a word list in which the word families also contain the specific derived forms that align with the *derivational morphology* included in the content for the reading exam for GCSE; DfE, 2022.) To the best of our knowledge, this is the only such detailed schedule available. Having it embedded within the profiler allows the tracking of systematic growth of lexical and grammatical systems over time and the creation of classroom resources compatible with that growth. For example, in the schedule, about one lesson per term is set aside for the study of 'rich texts'; also, throughout the schedule, many of the listening and reading texts relate to cultural, personal, social, or geographical themes (for examples, see https://ncelp.org/cultural-collection/). Thus, the profiler helps materials creators to know how much of a text is likely to be familiar to the learner (has been taught) and how much is unencountered language to date.

This profiling approach has limitations, of course. On the one hand, it overestimates knowledge, as learners are unlikely to reliably hold (even declarative) knowledge of words and grammar after having practiced them in just one lesson. Thus, the profiler only reflects the 'maximum' language introduced to date. On the other hand, it underestimates knowledge, as some learners could infer the meaning of words that have not been taught and understand words inflected with untaught inflections. Nevertheless, for our context, materials used by the teacher are likely to be a reasonable proxy for total exposure, in most cases.

Transfer-appropriate

The principle of transfer-appropriate practice can be interpreted and operationalised in several ways. Here we focus on the ideas that (1) declarative knowledge of vocabulary

grammar, and phonics, at least in the very early stages, often needs to be established in both written and oral modalities, with a view to applying that knowledge for oral and written receptive and productive skills (note that although declarative knowledge is often considered to be transferable, if a learner does not yet have reliable knowledge of sound-spelling relations, having knowledge of a written form only does not necessarily transfer to the oral modality); (2) that knowledge should be practiced in such a way that it can be proceduralized and automatized in different modalities and modes; and (3) practicing language in the kinds of context for which it is required (e.g., a translation, a role play, conversation, comprehension of texts) is likely to be helpful. These ideas are reflected in our materials in that each lesson involves introducing knowledge (lexical, morphosyntactic, sound-symbol correspondences) in different modes (comprehension and production) and modalities (oral and written). As *declarative* knowledge may be available for use in different modes or modalities (and particularly sound-spelling relations become more reliable), brief explanations of language patterns are given—in writing and also read out by the teacher—prior to practice in these different modes and modalities. Proceduralisation in different modes and modalities is supported as language content is practiced (in isolation and in context), in oral and written comprehension and production.

In order to operationalise transfer-appropriate practice, it can help to make features 'task essential', so that learners need to perceive or produce the feature to understand or create meaning (Loschky & Bley-Vroman, 1993). But achieving this requires specially designed activities particularly for features that are communicatively redundant, have low salience, or have complex L1-L2 relations. Trapping forms in *receptive* tasks is rather straightforward (VanPatten & Cadierno, 1993). However, for production, creating meaningful practice can be challenging. Lightbown (2000) defined practice as "opportunities for meaningful language use" (p. 443) and Lyster and Sato (2013) argue for "contextualized practice" (p. 83). But what does *meaningful* or *contextualized* practice look like for developing procedural knowledge to support communicative skills for such early stage learners? To address this need, we created information gaps for paired or small group oral production for every lesson (see Figure 4 for an example). These align with Gatbonton and Segalowitz's (2005) proposal that genuinely communicative practice requires information exchange for successful task completion (see also Sato & McDonough, 2019, an OASIS summary of which is used in our professional development materials). Each production task requires both the speaker(s) and the listener(s) to use the target feature to understand or produce meaning from the lexicon and/or from morphosyntax (e.g., the meaning of question words;

information about tense, number, person, gender, syntactic role [subject, object, agent, patient]; sentence type [declarative, negation, interrogative][6]).

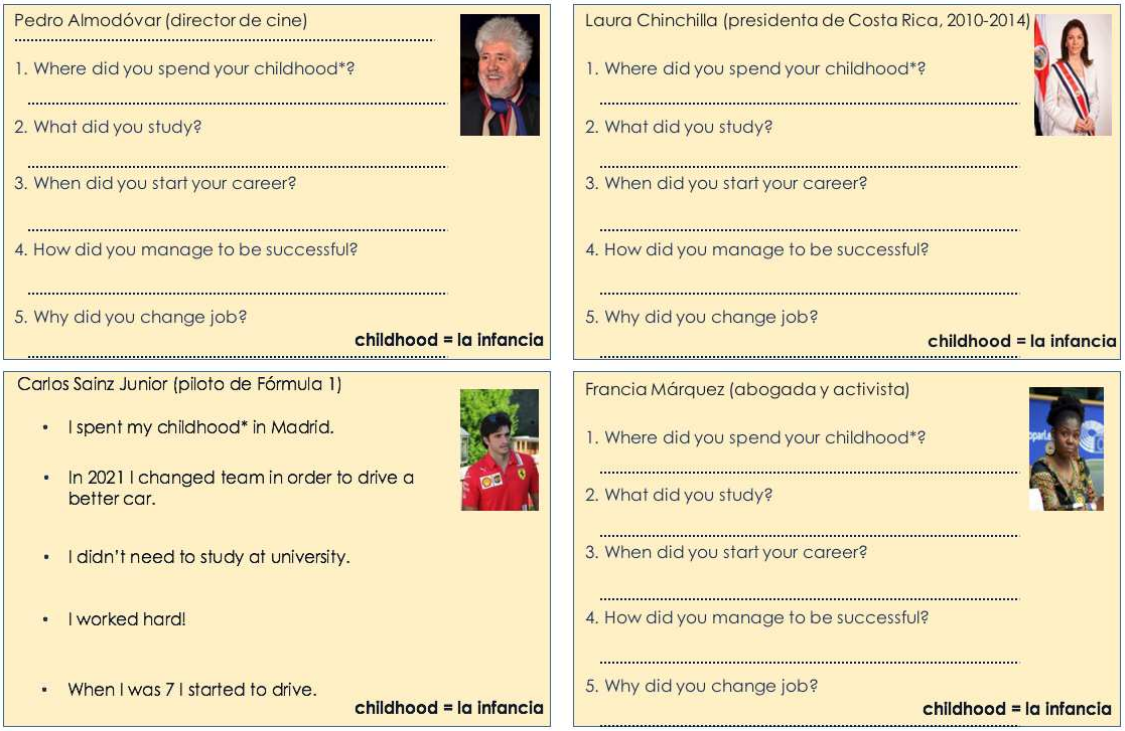

Figure 4: Extract of a four-person information gap from a lesson revisiting preterit interrogatives and declaratives in Spanish, in which each student has information about one character and asks questions about three others. Lexicon, including interrogative pronouns, are task-essential; preterit is task-likely or task-useful. For full lesson resource, see https://resources.ncelp.org/concern/resources/mk61rk014?locale=en.

Practice such as freer oral or written production is introduced slowly, and will be foregrounded more in the later stages of the curriculum.

Feedback

NCELP's perspective on feedback aligns with the notion that corrective feedback can help both to establish accurate and reliable declarative knowledge and to support proceduralization as knowledge about how to use the declarative knowledge becomes 'repackaged' in accessible chunks[7]. Proceduralization can be characterised by a gradual

[6] This metalanguage is not a feature to be learned, and is used here for presentational purposes.

[7] To enrich teacher understanding of the role of feedback in practice, our professional development events included hands-on activities to engage teachers' understanding of feedback types (e.g., recasts, prompts, metalinguistic), their effectiveness for learning in different contexts (e.g., error types or cognitive individual differences), and their role in extending motivational interactions.

reduction in errors, an increase in speed, and a decrease in the *variation* of speed (McManus & Marsden, 2019b). Many input-based whole class activities include referential style listening and reading activities from within processing instruction (Kasprowicz & Marsden, 2018; Marsden, 2006; Marsden & Chen, 2011). Such activities always include feedback built in as animations in the Power Point presentations, either after each of the initial few items with a view to establishing reliable declarative knowledge for all learners or at the end of the activity. Written and spoken production activities also have the correct responses provided so that learners receive feedback during or very shortly after the activity. Given that several studies have pointed to the benefits of immediate over delayed feedback (e.g., Arroyo & Yilmaz, 2018; Fu, 2022), computer assisted language learning tools (for vocabulary, Quizlet; for grammar, Gaming Grammar https://www.gaminggrammar.com/ used by Kasprowicz, Marsden, & Sephton, 2020) are weaved into the practice schedule (though see Li, 2020 and Kim & Webb, 2022 for debate about the benefits of immediate versus delayed feedback for different types of treatments and practice schedules).

Challenging (desirable difficulty)

Incorporating the principle of desirable difficulty at a classroom level is a challenge, given its sparse and context-relevant evidence base and interactions between difficulty and individual differences. Nevertheless, our curriculum and resources and professional development activities have emphasized 'challenge'—broadly defined—as an important principle of practice, operationalized in various ways, of which we give several examples here.

- Learners cannot rely on guess work (for example, they could not use only pictures or world knowledge to answer questions, a common characteristic of exercises in some textbooks; see Marsden, 2005).
- Form has to be processed (from input) or produced (in output) to get meaning or function.
- Spacing between sets of words and grammatical and sound subsystems was implemented in expanding practice schedules.
- Interleaving has been used at various different levels of language representation (phonemic, morphological, lexical, and syntactic), modes (receptive, productive), modalities (oral, written), and at different levels of distribution (within a single activity, within a lesson, and across weeks, months, and even years), leveraging both spacing and interleaving[8].

[8] For example, referential style input-based activities juxtapose different sets of form-meaning mappings,

- Our cumulative approach to lexical and morphosyntactic development and testing requires active recall of earlier material.
- Cognitive load is manipulated by different task characteristics varying parameters such as time available (timed versus untimed), familiarity (with and without repetition), communicative pressure (monologic versus dialogic; whole class vs individual) and (bi-)modality (with quasi-synchronized text + oral input aiding segmentation and comprehension; Peters & Muñoz, 2020).
- Language-focused activities that draw on knowledge in different modalities, such as dictation and translation across modalities (e.g., L1 written to L2 oral translation; L2 written to L1 oral translation) provide opportunities for controlled access to declarative knowledge, as well as for proceduralisation and automatization.
- During paired oral production activities, access to written resources that can provide the precise language necessary to complete the task is not encouraged, so that learners must actively retrieve the language needed.

More generally, theorizing about the notion of challenge is broad, and we drew on a range of frameworks. For example, dictoglosses are used to provide opportunities for 'pushed output' (Swain, 2005) and increasing 'involvement' is reflected in activities that require 'search', 'need', or 'evaluation' of meaningful content in texts (see Yanagisawa & Webb's 2021 meta-analysis). Oral production tasks that impose communicative and/or time pressure will occur more often and in more open-ended formats in materials for years 10 and 11 (14-16-year-olds). We emphasise, however, that whilst ideas for additional support or extension activities are sometimes provided in notes accompanying the materials, teachers' judgement is required to adapt materials to suit their learners.

Insights into automatization and explicit-implicit perspective

Given the learners' limited exposure and the wide range of cognitive abilities in this population, it is a challenge to know what the right balance is between explicit information and mechanical practice to establish declarative knowledge versus opportunities to establish and automatize proceduralized (chunked) knowledge. Whether

and later recycle them in juxtaposition with others mappings (see McManus & Marsden, 2019a, for an example in which French imperfect endings were juxtaposed with present tense in some activities and perfect tense in others).

the ultimate purpose was automatized knowledge or implicit knowledge (without awareness) was, in fact, of little relevance to our stakeholders; and, in any case, we had no way of testing knowledge types in the classroom context in the absence of sophisticated equipment to measure reaction times or eye movements or the capacity to elicit subjective judgements of awareness. For pedagogical purposes, the aim is to facilitate use of knowledge that has the hallmarks characteristic of automatized knowledge: accessed fast (without a lot of pre- or during-task planning); reliably; with decreasing variance in the speed of responses (McManus & Marsden, 2019b); and would not readily fade. We took several research-informed steps to promote the development of automatized explicit knowledge. For example, time pressure (such as L2-L1 and L1-L2 meaning recall of vocabulary against a 30 second timer) and repetition of comprehension and production activities are frequently used to support proceduralisation and automatization.

One way in which we had to consider relations between different knowledge types was in the design of our assessments. For each year of the course, we provide two sets of *achievement tests,* and one set of *applying your knowledge tests*. The *achievement tests* assess knowledge via closed-option tests of isolated items of vocabulary, grammar, sound-symbol correspondences. The rationale for assessing these isolated components drew on the vast bodies of research demonstrating how lexical, grammatical, phonological and orthographical knowledge strongly predict reading, listening, speaking, and writing proficiency (see Jeon & In'nami, 2022, for syntheses). The *applying your knowledge* tests (towards the end of each school year) sample less systematically from the whole body of language taught to date as they are more akin to proficiency tests. Several design features of the tests reflect characteristics of tasks that are likely to elicit, at least in part, automatized (procedural) knowledge and/or implicit knowledge. For example, test items are at sentence and paragraph level (rather than at the phoneme, grapheme, morpheme, or word levels), and understanding is assessed by comprehension questions or by learners extracting ideational units. Little preparation time is given for production tasks, and marks are awarded primarily for communication/understanding of meaning. The speaking, writing, reading, and listening tests allow demonstration of knowledge that is accessible across different modes and modalities. Inaccuracies are tolerated for certain features, to recognize that proceduralized knowledge is prone to inaccuracies when pressures are exerted on memory systems.

As noted above, the whole pedagogy is not solely based upon deliberate practice; the resources also provide opportunities to implicitly tally already-noticed language and/or engage in inductive and/or incidental learning. These could *potentially* feed implicit learning and serve to establish implicit knowledge. However, with about two

hours of weekly exposure to the language, totaling 400-450 hours, automatized knowledge and/or implicit knowledge is unlikely to be achieved by all learners for all of the content covered. To date, we do not have data on the extent to which this practice-based curriculum has succeeded in automatizing knowledge.

Concluding remarks: Ways forward for research.

We have illustrated how our curriculum and pedagogy draws upon or reflects principles of practice, but we now conclude by indicating several areas in which we found research was either lacking in relevance or absent.

Above we illustrated how practice schedules at the scale of a whole curriculum require expectations about linguistic development to be made explicit. For practice schedules to bear fruit, they require patience for learners to establish reliable declarative and proceduralized knowledge before automatization is attained. Thus, an approach that is driven by practice of components of language knowledge must go hand in hand with (increased) tolerance for the dynamic and errorful nature of development, where accuracy, fluency, and complexity can vary when new language is introduced into the developing system (e.g., for writing, Huang, Steinkrauss, & Verspoor, 2021). This need for patience is in stark relief when compared to heavily formulaic-based topic-driven approaches, where rote-learned chunks can lend a veneer of efficiency and fluency to oral and written production at very early stages. Whilst such formulae *may* provide a model for implicit (statistical) learning or incidental analysis, these chunks are not reliably or appropriately analysed to lead to meaningful manipulation of language by all learners in the early stages, at least not in this kind of low exposure context (e.g., Myles, Hooper, & Mitchell, 1998). It remains to be investigated whether alternative practice schedules could usefully incorporate more formulaic chunks and yet (a) still retain systematic and frequent revisiting of language (vocabulary, grammar, and sound-spelling relations) and (b) not demotivate learners by repeating the same topics in which those formulaic chunks are used (e.g., daily routine for practising reflexives). This is an important question for future research.

In addition to collecting feedback provided by teachers, there is an acute need to formally evaluate this language-practice based curriculum. But evaluation of effectiveness is particularly challenging for *whole* curriculum and pedagogy projects (indeed, Collins & Muñoz's 2016 review found these to be lacking in number). With so many variables changing, research-informed principles cannot be isolated to one or two components. Thus, the 'grain' size of what should be evaluated is not clear. Should we evaluate, in tightly controlled experiments, the specific approach to input-based

grammar practice, isolated phonics practice, spacing in our vocabulary practice schedules, the approach to cumulative testing—that is, the kind of Instructed SLA research that is often published in journals? Or should we somehow compare 'everything' to something else in its entirety?—but what? We call for robust and multi-faceted longitudinal evaluations of holistic curriculum and pedagogy packages. But we also emphasise the need to continue the current practice research agenda: a piecemeal and componential approach to improving our understanding of practice, in both classroom and laboratory settings.

Indeed, many such agendas seem to be wide open. For example, is phonics practice most effectively massed at the start of a programme, interspersed in shorter isolated bursts (as we incorporated), or integrated within other content in a more reactive form-focused approach? There is certainly a need for more research into *expanding* practice schedules and *cumulative* testing. In particular, we need research that manipulates the 'unit' for interleaving and yet is also ecologically valid. In reality, some kind of interleaving is surely inevitable because no scheme of work could purely mass practice every lesson. Doesn't *all* instruction interleave content at different grain sizes: phoneme, morpheme, word, syntax? In our expanding practice schedules, interleaving had to happen for several different levels simultaneously yet with each level being on different practice schedules. For example, one feature could be interleaved at a micro level *within* one activity (e.g., contrasting the function of '-o' with –'a' verb inflections for conveying person in the present tense in Spanish); one of those same grammar features would then reappear several weeks later but within a different micro-interleaving pair (e.g., '-o' versus '-é' for conveying tense in the first person). In contrast, a lexical set could be interleaved with other sets over a longer time frame, over several weeks or even months; and *within* the activities and lesson itself, items in that one vocabulary set could be distributed at different intervals ranging from a few seconds to several minutes. The spacing of different granularities of the language to be inter-leaved is in reality very difficult if not impossible to control tightly.

Another body of research that we could not draw upon in a meticulous way was that on optimal inter-study interval (ISI: spacing between practice opportunities). The existing evidence focuses on intervals of minutes, hours, or days—which could provide only broad heuristics for curriculum design and pedagogy, given that the times of lessons are fixed by individual school timetables. But we needed evidence about intervals of weeks, months, and years. In the end, the information we had to determine such intervals was (a) the time available, each week and in total, and (b) the approximate number of words to be learned in total. Similarly, the idea that an optimum ISI is determined by the retention interval (RI: the time between the last learning

opportunity and the test)–the so-called ISI-RI ratio–did not seem directly relevant to the design of a large, 5-year curriculum that culminates in an external examination, because research in this area has not yet, to our knowledge, used a 'final retention test' that is actually a high-stakes proficiency test administered at a time that is beyond the instructors' control. So, this research has not worked with 'real-world' bodies of language knowledge. Language proficiency tests—a real-world analogy for the laboratory's final 'retention tests'— cover vast amounts of knowledge and skills, so a researcher can neither pinpoint a specific retention interval (as one cannot identify the distance between a single language feature that is tested in a proficiency test and when it was last practiced) nor the interval between practicing a relevant language feature (as the tests are holistic, rather than testing isolated bodies of knowledge). Even at a more fine-grained level of planning during the years prior to the high-stakes external tests, the internal school tests are administered at times that are almost always beyond individual instructors' control. Compounding the problem is that even these assessments need to test mixed bodies of knowledge, the subcomponents of which would each have been taught at different times and at different intervals. Thus, it was futile to advise a specific ISI-RI ratio, as this would have been impossible to make a reality for all of the language content being tested in any one lesson.

Similarly, as discussed above, the notion of an expanding practice schedule had to be interpreted loosely for several reasons. One was that research is still needed on the benefits of uniform, expanding, or contracting schedules, for different language content (e.g., vocabulary, grammar), for different pedagogical aims (e.g., skills or knowledge) and treatments (e.g., feedback, input-based), and with individual differences in mind (e.g., ages, proficiencies, motivation types, working memory capacity). Another perhaps more important reason was that adhering inflexibly to any schedule (uniform, expanding, or contracting) could not be operationalised in reality at either macro (between session) or micro (within session) levels without compromising other aspects of the curriculum and pedagogy.

This does not detract from the value in pursuing these avenues of research to improve our understanding of human cognition and language learning. Our point is that the current state of the science could not inform to a very fine degree how we addressed these particular problems. Critically for future work in this area, our practice schedules are fully adaptable. Others are welcome to develop different versions that, for example, strive to space words more closely (say, every 2 and 5 weeks; or uniformly, every 2 and 4 weeks) and then try to weave the other language elements, such as the grammar and sound-writing systems, around that spacing of the lexicon. To inform such planning, we created a 'vocabulary spacing calculator' in excel, in which a user can manipulate

certain parameters, such as how many words are introduced each lesson, how many words are to be learned in total, how many weeks there are between each set of words, how many intentional revisits of each word are incorporated, and how many bulk revision lessons would be appropriate before a test or an exam. This (in progress) calculator is available upon request, though it has limitations, such as it aggregates across series of weeks or months (rather than allowing full flexibility in terms of numbers of words in different sessions), and it only supports curriculum planning in terms of the lexicon.

Another gap in our evidence base was specific recommendations about how individual differences interact with pedagogical decisions. Some research, albeit in need of replication with more robust measures, suggests that brief deductive presentation of information is more likely to help more learners most of the time than more inductive approaches (e.g., Erlam, 2005), especially given that analytic ability can determine the effectiveness of practice (e.g., Kasprowicz et al. 2019). However, operationalising any findings from this aptitude*treatment agenda poses challenges for classrooms with a range of aptitudes and cognitive capacities. We recommend that such research needs to be firmly grounded in treatment types (e.g., computer-assisted language learning tools) that can feasibly vary an intervention within a single classroom.

The overarching lesson we have learned is that before embarking on an investigation into practice in any one context, it is necessary to define the content. Whilst this is clearly critical for educators, testers, and researchers alike, we found that identifying the 'what' of practice was highly politically sensitive. Researchers' tendency to work in specific subdomains may have meant that we have not given sufficient attention to defining the 'what'. To caricature a little: The vocabulary and corpus-based researchers identify useful word lists; the interventionists (task-based, input-based, error-correction) identify effective treatments; linguistically-minded researchers identify routes and rates of acquisition and their different predictors; test-oriented researchers define constructs of proficiency and competence and their correlates; and practice-schedule researchers examine intricacies of different spacing and retention intervals for pre-determined and relatively small bodies of language. Although we harnessed findings from these separate domains, we could do nothing practical to help teachers until we had first defined 'what' should be learned and when.

To define the content, we had definitions of language competence from the 1980s and 1990s and different types of syllabi (synthetic [componential] through functional-notional to task- and content-based approaches) each with different implications for 'what' should be practiced. The less defined the body of knowledge is, the more it may be beneficial to learn strategies to cope when comprehension or

production breaks down (see Graham et al., 2020 for work in this area). In such a context, some of the 'what' should be practiced would then need to be conceived of as declarative knowledge of strategies (e.g., 'to guess the meaning of an unknown word, I know that I need to use other words in the surrounding context' or 'if I can't say exactly what I want, I know that I need to use other ways of expressing my meaning'). These strategies may then become proceduralized via practice in carrying out the strategies in a conscious way, and then perhaps automatized so that lexical inferencing or circumlocution occur reliably without awareness or effort. Indeed, in our resources, we incorporated some practice in lexical inferencing skills, particularly as inferencing also helps to consolidate the meaning of the *familiar* words. However, on the whole, we have mainly focused so far on identifying and establishing knowledge and automatizations of forms, functions, meanings of *language* itself, rather than strategies. This was because effective strategy use generally seems to be better learned and automatized among post-beginner, intermediate, and advanced learners than beginner learners (Plonsky, 2011). Thus, at later stages of the curriculum, we will perhaps include some strategy instruction, e.g., discourse planning in writing and circumlocution in speaking. Research is certainly needed to investigate effective balances in the earliest stages of learning in contexts with limited curriculum time of establishing and automatizing the knowledge and use of language itself versus knowledge and use of strategies.

It seems that identifying and sequencing ecologically valid bodies of knowledge is critical if findings from practice-based research are to stand a chance of influencing main stream instruction at a programme level. So, we end by recommending that some researchers investigating practice consider bodies of knowledge at a less fine-grained level than hitherto and work with teachers, test-developers, and policy-makers who have to see 'curricula' more holistically, spanning months and years. If we do not do this, we run the risk that the practice research agenda will be lost on educators, materials writers, and policy-makers.

References

Adesope, O. O., T. Lavin, T. Thompson, & C. Ungerleider (2011). Pedagogical strategies for teaching literacy to ESL immigrant students: a meta-analysis. *British Journal of Educational Psychology 81* (4), 629–653.

Arroyo, D. C. & Yilmaz, Y. (2018). An open for replication study: The role of feedback timing in synchronous computer-mediated communication. *Language Learning 68*(4), 1–31.

Brooks, P. J., Kempe, V., & Sionov, A. (2006). The role of learner and input variables in learning inflectional morphology. *Applied Psycholinguistics*, *27*(2), 185–209.

Brumfit, C. J. (1995). Teacher professionalism and research. In G. Cook and B. Seidlhofer (Eds): *Principle and Practice in Applied Linguistics* Oxford: Oxford University Press, pp. 27–41.

Collins, L., & Marsden, E. (2016). Cognitive perspectives on classroom language learning. In G.Hall (Ed.), *The Routledge Handbook of English Language Teaching* [Chapter 21] (Routledge Handbooks in Applied Linguistics). Routledge.

Collins, L., & Muñoz, C. (2016). The foreign language classroom: Current perspectives and future considerations. *The Modern Language Journal*, *100(1),* 133–147.

Department for Education. (2015). *Modern foreign language: GCSE subject content*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/485567/GCSE_subject_content_modern_foreign_langs.pdf

Department for Education. (2022). *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

DeKeyser, R. (2010). Practice for second language learning: Don't throw out the baby with the bathwater. *International Journal of English Studies, 10*(1), 155–165.

DeKeyser, R. (2016). Of moving targets and chameleons: Why the concept of difficulty is so hard to pin down. *Studies in Second Language Acquisition, 38*(2), 353–363.

DeKeyser, R. M. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15–32). New York, NY: Routledge.

Dudley, A., & Marsden, E. (under revision). Exploring the lexical content of the listening and reading tests in the GCSE exams in French, German, and Spanish in England. *Language Testing.*

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*, 143–188

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research, 9* (2) 147 – 171.

Finlayson, N., & Marsden, E. (2022). Assessing depth of word knowledge of beginner learners of French, German, and Spanish aged 11-14 in England. In S. W. Chong, & H. Reinders (Eds.), *Learner-Centred Approaches to Language Assessment* Palgrave Macmillan.

Finlayson, N., Marsden, E., Hawkes, R., & Avery, N. (in progress). General service word lists for beginner and low-intermediate learners of French, German, and Spanish.

Finlayson, N., Marsden, E., Anthony, L., Bovolenta, G., & Hawkes, R. (2021). *MultiLingProfiler (Version 2) [Computer software]*. University of York. https://www.multilingprofiler.net/

Fu, M. (2022). The effects of immediate and delayed corrective feedback on 2 development. *Studies in Second Language Acquisition, 44*(1), 2-34.

Gerbier, E., Toppino, T., & Koenig, O. (2015). Optimising retention through multiple study opportunities over days: The benefit of an expanding schedule of repetitions. *Memory, 23*(6), 943-954.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Gor, K. (2010). Introduction. Beyond the obvious: do second language learners process inflectional morphology? *Language Learning, 60*, 1–20.

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. *The Modern Language Journal, 104*, 693–714.

Häcker, M. (2008). Eleven pets and 20 ways to express one's opinion: The vocabulary learners of German acquire at English secondary schools. *Language Learning Journal*, *36*(2), 215–226.

Harrison, S., Glover, S., Dzuilka-Heywood, C., Moores, A., & Gilles, J.-C. (2016). *AQA GCSE French Foundation*. Oxford University Press.

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

Horvath, J. C., Lodge, J. M., & Hattie, J. (2016). *From the laboratory to the classroom: Translating science of learning for teachers*. New York, NY: Routledge.

Housen A, De Clercq B, Kuiken F, Vedder I. (2019). Multiple approaches to complexity in second language research. *Second Language Research.* 35(1):3-21.

Huang, T. Steinkrauss, R., & Verspoor, M. (2021). Variability as predictor in L2 writing proficiency. *Journal of Second Language Writing, 52,* 100787.

Jeon, E. H., & In'nami, Y. (2022). *Understanding L2 proficiency: Theoretical and meta-analytic investigations*. Amsterdam, the Netherlands: John Benjamins.

Kasprowicz, R. E., & Marsden, E. (2018). Towards ecological validity in research into input-based practice: Form spotting can be as beneficial as form-meaning practice. *Applied Linguistics*, *39*(6), 886–911.

Kasprowicz, R. E., Marsden, E. J., & Sephton, N. (2019). Investigating distribution of practice effects for the learning of foreign language verb morphology in the young learner classroom. *The Modern Language Journal*, *103*(3), 580-606.

Keßler, J.-U., Lenzing, A., & Liebner, M. (Eds.). (2016). *Developing, modelling, and assessing second languages*. Amsterdam: John Benjamins.

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.

Kim, Y., Kang, S., Yun, H., Kim, B., & Choi, B. (2020). The role of task repetition in a Korean as a foreign language classroom: Writing quality, attention to form, and learning of Korean grammar. *Foreign Language Annals*, *53,* 827-849.

Kim, S., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning, 72,* 269–319.

Li, S. (2020). What is the ideal time to provide corrective feedback? Replication of Li, Zhu & Ellis (2016) and Arroyo & Yilmaz (2018). *Language Teaching, 53,* 96–108.

Lightbown, P. (2000). Classroom SLA research and second language teaching. *Applied Linguistics, 21*(4), 431–462.

Loschky, L., & Bley-Vroman, R. (1993) 'Grammar and task-based methodology', in G. Crookes and S. Gass (eds) Tasks and language learning: Integrating theory and practice. Clevedon, England: Multilingual Matters Ltd. 123–167.

Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In M. P. García Mayo, J. Gutierrez-Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 71–92). John Benjamins.

Marsden, E. (2005). Input-based grammar pedagogy: a comparison of two possibilities. *Language Learning Journal*, *31*(1), 9-20.

Marsden, E. (2006). Exploring input processing in the classroom: An experimental comparison of processing instruction and enriched input. *Language Learning*, *56*(3), 507-566.

Marsden, E., Alferink, I., Andringa, S., Bolibaugh, C., Collins, L, Jackson, C., Kasprowicz, R., O'Reilly, D., Plonsky, L. (2018). Open Accessible Summaries in Language Studies (OASIS) [Database]. Available at https://www.oasis-database.org

Marsden, E. J., & Chen, H-Y. (2011). The roles of structured input activities in processing instruction and the kinds of knowledge they promote. *Language Learning*, *61*(4), 1058–1098.

Marsden, E., & David, A. (2008). Vocabulary use during conversation: a cross-sectional study of development from year 9 to year 13 among learners of Spanish and French. *Language Learning Journal*, *36*(2), 181-198.

Marsden, E, Dudley, A., & Hawkes, R. (under revision). Use of word lists in a high-stakes, low-exposure context: Topic-driven or frequency-informed. *The Modern Language Journal.*

McManus, K., & Marsden, E. J. (2019a). Using explicit instruction about L1 to reduce crosslinguistic effects in L2 grammar learning: Evidence from oral production in L2 French. *The Modern Language Journal*, *103*(2), 459-480.

McManus, K., & Marsden, E. J. (2019b). Signatures of automaticity during practice: Explicit instruction about L1 processing routines can improve L2 grammatical processing. *Applied Psycholinguistics*, *40*(1), 205-234.

Mills, N., Pajares, F., & Herron, C. (2007). Self-efficacy of college intermediate French students: Relation to achievement and motivation. *Language Learning, 57*, 417 – 442

Milton, J. (2006). Language Lite? Learning French vocabulary in school. *French Language Studies, 16*(2), 187–205.

Milton, J. (2015). French lexis and formal exams in the British foreign language classroom. *Revue française de linguistique appliquée* 1, 107-119

Milton, J. & Meara, P. (1998). Are the British really bad at learning foreign languages? *Language Learning Journal*, 18: 68–76.

Mitchell, R. (2002). Foreign language education in an age of global English. *Centre for Language in Education Occasional Papers 57.* University of Southampton

Myles, F., Hooper, J. & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning, 48* (3), 323-363.

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition, 37*(4), 677-711.

Nakata, T. & Suzuki, Y. (2019). Mixing grammar exercises facilitates long-term retention: Effects of blocking, interleaving, and increasing practice. *The Modern Language Journal,* 103, 629-647.

Nakata, T., Tada, S., McLean, S., & Kim, Y. (2021). Effects of distributed retrieval practice over a semester: Cumulative tests as a way to facilitate second language vocabulary learning. *TESOL Quarterly,* 55, 248-270.

Peters, E. & Muñoz, C. (2020). Introduction to the special issue: Language learning from multimodal input. *Studies in Second Language Acquisition, 42*(3), 489-497.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: a meta-analysis. *Language Learning, 61*, 4, 993–1038.

Sato, M., & McDonough, K. (2019). Practice is important but how about its quality? Contextualized practice in the classroom. *Studies in Second Language Acquisition, 41*(5), 999-1026.

Schmitt, N. (2010). *Researching vocabulary*. Palgrave Macmillan UK.

Schuetze, U. (2015). Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long-term memory. *Language Teaching Research*, *19*(1), 28-42.

Schuetze, U., & Weimer-Stuckmann, G. (2011). Retention in SLA lexical processing. *CALICO Journal*, *28*(2), 460-472.

Suzuki, Y. (2017). The Optimal Distribution of Practice for the Acquisition of L2 Morphology: A Conceptual Replication and Extension. Language Learning, 67, 512-545.

Suzuki, Y. (2021). Optimizing fluency training for speaking skills transfer: comparing the effects of blocked and interleaved task repetition. *Language Learning, 71*, 285-325.

Suzuki, Y., Nakata, T. and DeKeyser, R. (2019). Optimizing Second Language Practice in the Classroom: Perspectives from Cognitive Psychology. *The Modern Language Journal,* 103, 551-561.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–484). Mahwah, NJ: Lawrence Erlbaum.

Teaching Schools Council, 2016 *Modern Foreign Languages Pedagogy Review*. *Teaching Schools Council. https://ncelp.org/wp-content/uploads/2020/02/MFL_Pedagogy_Review_Report_TSC_PUBLISHED_VERSION_Nov_2016_1_.pdf*

Tschichold, C. (2012). French vocabulary in Encore Tricolore: Do pupils have a chance? *Language Learning Journal*, *40*(1), 7–19.

Uchihara T, Eguchi M, Clenton J, Kyle K, Saito K. (2022) To what extent is collocation knowledge associated with oral proficiency? A corpus-based approach to word association. *Language and Speech*, *65*(2), 311-336.

VanPatten, B., & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in Second Language Acquisition, 15*(2), 225-243.

Wallace, M.P. (2022). Individual differences in second language listening: examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, *72*, 5-44.

Woore, R. (2011). Investigating and developing beginner learners' decoding proficiency in second language French: an evaluation of two programmes of instruction. Ph.D. thesis, University of Oxford

Woore, R., S. Graham, A. Porter, L. Courtney, & C. Savory (2018). Foreign language education: unlocking reading (FLEUR) - A study into the teaching of reading to beginner learners of French in secondary school.

Yanagisawa, A., & Webb, S. (2021). To what extent does the Involvement Load Hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning, 71,* 487–536.