Diverse population, homogenous ability: the development of a new receptive vocabulary size test for young language learners in England using Rasch analysis

Research Methods in Applied Linguistics 3 (2024) 100166

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Diverse population, homogenous ability: The development of a new receptive vocabulary size test for young language learners in England using Rasch analysis

Nicola Morea [a,*], Rowena Eloise Kasprowicz [a], Astrid Morrison [b], Carmen Silvestri [a]

[a] University of Reading, United Kingdom
[b] Universidad Autónoma de Chile, Chile

ARTICLE INFO

Keywords:
Young language learners
Vocabulary
Language testing
Longitudinal
Rasch analysis

ABSTRACT

With mandatory second language learning in primary education becoming the norm worldwide, research investigating young language learners' (YLLs) linguistic development has increased. However, designing language tests appropriate for YLLs poses unique challenges due to population characteristics and variability in national and institutional contexts. In this article, we present a new vocabulary test designed to track the rate of progression in receptive vocabulary size of primary school children learning French, German or Spanish in England.

Test content was selected after an analysis of programmes of study commonly used in primary schools in England. The test required two validation phases using Rasch analysis. The initial tests were administered to 1662 students from Year 3 (7–8 years old) to Year 5 (9–10 years old). All tests showed poor person reliability, which was driven by a mismatch between item difficulties and participant abilities. Various actions were taken in relation to vocabulary identification, test format and length, and sampling procedures, and the revised tests were re-administered to 2202 students from Year 3 to Year 6 the following year. As a result, person reliability considerably improved, and all test versions showed good fit to the Rasch model.

Drawing on the lessons learnt, we discuss some of the key population- and context-related challenges of designing robust language tests for beginner YLLs learning a language other than English in input-poor, instructed contexts. Further, we provide recommendations on suitable approaches for test-item identification, test format and length, and data analysis.

With compulsory second language (L2) learning in primary schools becoming the norm worldwide (Nikolov & Timpe-Laughlin, 2021), research on young language learners' (YLLs) progression is particularly important to evaluate the effectiveness of educational policies, as well as understand the extent to which YLLs develop language skills in instructed contexts and the rate at which they do so. However, language assessment research on YLLs poses unique methodological challenges related to both population characteristics and the specific national and institutional contexts in which language learning takes place. YLLs differ from older and adult learners in various aspects, such as having a shorter attention span and varying degree of literacy in the language of instruction (Bailey, Heritage, & Butler, 2014). Furthermore, YLLs may have limited assessment literacy, namely an understanding of assessment procedures and their meaning (Weng & Liu, 2024). Additional contextual issues stem from the variability in the language provision for

* Corresponding author at: University of Reading, Cambridge, United Kingdom.
E-mail address: n.morea@reading.ac.uk (N. Morea).

https://doi.org/10.1016/j.rmal.2024.100166
Received 15 May 2024; Received in revised form 30 October 2024; Accepted 1 November 2024
Available online 12 November 2024

YLLs across national contexts, such as amount of instructed time, curriculum content and objectives, assessment procedures, and teacher preparation, which may complicate the choice of what language content YLLs should be assessed on.

In contexts presenting huge variation in population characteristics and language teaching provision, the design of reliable and valid language tests seems particularly challenging, and yet necessary. Among the various aspects of language that can be assessed, vocabulary knowledge represents a foundational element in early language learning (Butler, 2019) and a core aspect underpinning multiple language skills, such as listening, reading and writing (Stæhr, 2008). Accordingly, this article presents the design and validation of a novel test of receptive vocabulary size for primary school children in England aged 7–11. Specifically, we will showcase how challenges related to (i) YLLs' characteristics, (ii) the English national context, and (iii) the longitudinal research design were addressed during the two phases of test development. Additionally, we will evaluate the opportunities that Rasch analysis provides for designing and validating research instruments to longitudinally track YLLs' L2 progression.

1. Literature review

1.1. Young language learners: definitions, characteristics, and the English context

In this article, we follow Hasselgren's (2012) definition of YLLs as "primary school pupils up to about 12 years who are learning a second, additional or foreign language" (p. 93). YLLs have specific characteristics that distinguish them from older and adult learners. Compared to older/adult learners, YLLs have a shorter attention span and a slower processing speed (Bailey et al., 2014; McKay, 2006), resulting in the need to develop tests and instruments that are concrete rather than abstract, that reflect the lived experiences of YLLs and that are of appropriate length so as to sustain learners' motivation (Courtney & Graham, 2019; McKay, 2006). Additionally, YLLs may be linguistically diverse; for example, in 2023 in England, 22 % of children were believed to use a language other than English at home (DfE, 2024). As a result, variability in YLLs' literacy and proficiency in the language of instruction should also be considered when developing research instruments. Given these considerations, research methods designed for older learners may not be appropriate for YLLs, requiring the design of instruments specifically tailored to this population (Kasprowicz, Graham, & Morea, forthcoming). Moreover, whilst research on YLLs tends to focus on the learning of L2 English (Nikolov & Timpe-Laughlin, 2021), there is a lack of instruments assessing YLLs' knowledge of languages other than English (LOTE) in instructed contexts with limited language input.

Population-specific characteristics are not the only aspects that researchers should consider when designing language assessments for YLLs, as the national and institutional context in which language learning occurs may pose additional challenges (Kasprowicz et al., forthcoming). For example, instructed L2 teaching in primary school settings is often characterised by limited L2 input and variability in curriculum content, as is the case in England, the context of this study. Here, foreign languages have been a compulsory subject in primary education since September 2014, with the clear expectation that learners should make "substantial progress in one language" by the end of primary school (DfE, 2013, p. 2). However, the National Curriculum does not provide guidance on the language content that should be taught in primary schools, nor are there nationally agreed assessment criteria to evaluate whether the National Curriculum objectives are being achieved (McLachlan, 2009). Furthermore, limited time available for language learning (typically 30–60 min per week) (Collen, 2022) and variability in teacher language proficiency (Graham, Courtney, Marinis, & Tonkyn, 2017) complicate the picture. As a result, there is a clear need to understand the nature and rate of development of primary school children's linguistic knowledge in this context. However, the lack of a shared programme of study or assessment framework makes it particularly difficult to identify test content when assessing language learning across multiple schools (see Section 2.1 for further discussion).

1.2. Assessing vocabulary knowledge

Vocabulary knowledge has been defined as a "foundational element for language development" (Butler, 2019, p. 4), "one of the building blocks of language" and "basics of communication" (David, 2008, p.167). It is agreed that vocabulary knowledge underpins other language skills. In a meta-analysis reviewing findings from over 100 studies investigating the relationship between vocabulary knowledge in a second language (L2) and L2 reading and listening comprehension skills, Zhang and Zhang (2022) found that vocabulary knowledge accounted for a considerable amount of variance (31 %−45 %) in L2 reading comprehension. Similarly, research has also pointed to the strong association between vocabulary knowledge and L2 writing (e.g., Dabbagh & Janebi Enayat, 2019; Stæhr, 2008) and speaking ability (e.g., Koizumi & In'nami, 2013; Tong, Hasim, & Halim, 2022).

The construct of vocabulary knowledge is however multifaceted, and there does not appear to be a consensus in defining its components (Pignot-Shahov, 2012). Two conceptual distinctions are commonly made when investigating L2 vocabulary knowledge, namely between receptive and productive vocabulary (i.e., passive vs active knowledge) and between size (or breadth) and depth of vocabulary (i.e., number of known words vs how well words are known) (Edmonds, Clenton, & Elmetaher, 2022). In the context of testing the L2 vocabulary knowledge of YLLs at the early stages of L2 learning, focusing on L2 receptive vocabulary size was deemed most appropriate for this study, as productive knowledge and vocabulary depth require more time to develop (Laufer & Paribakht, 2008).

However, designing tests to assess YLLs' receptive vocabulary size poses certain challenges. Firstly, since a large number of items is needed in order to predict language learners' vocabulary size (Laufer, Elder, Hill, & Congdon, 2004), these tests tend to be relatively long and repetitive, a characteristic in contrast with the need for relatively short instruments to account for YLLs' attention span and desire for short and engaging learning activities.

Secondly, test content must be sampled from a large list of words or corpus reflecting the language test takers have been exposed to.

In the context of YLLs in England, this task is complicated by variation in curriculum content and the lack of corpora in French, German and Spanish capturing the target language students are exposed to in primary schools in England (see Section 2.1). Additionally, existing tests of vocabulary knowledge may be unsuitable for this context as they are designed for older learners and mainly rely on a word frequency approach to vocabulary sampling (Dudley, Marsden, & Bovolenta, 2024). Word frequency represents a common approach for selecting language to include in vocabulary tests, as it is assumed that high frequency words are more likely to be prioritised by teachers and appear in learning materials (Milton, 2008). However, as De Wilde (2023) argues, "frequency lists might be useful when deciding which words to teach but they might be less useful when trying to estimate learners' word knowledge" (p. 6), as young learners' vocabulary acquisition is influenced by a range of factors other than frequency, such as concreteness and cognateness. In the context of early L2 learning in instructed settings, it is indeed common for low-frequency vocabulary to be taught, and particularly cognates and thematic words (e.g., animals, colours, classroom objects) (Bardel, Gudmundson, & Lindqvist, 2012). Therefore, an approach for vocabulary sampling that accounts for both word frequency and curriculum content in the research context may be more suitable than a frequency-based approach alone (Dudley et al., 2024).

Finally, when the research aim is to track linguistic progression over multiple years, additional challenges emerge related to the longitudinal research design. These are discussed in the next section.

1.3. Longitudinal research designs

Longitudinal research refers to research designs conducted over time and involving multiple data collection points. Longitudinal studies that follow the same participants over time are considered more powerful than cross-sectional research to evaluate change or stability in a sample in relation to one or more variables of interest (Cohen, Manion, & Morrison, 2007). However, there are a number of factors to consider in longitudinal research design in order to ensure the validity of results.

In longitudinal assessment research, a central issue is the comparability of assessment data collected at two or more timepoints. Although the issue could be solved by administering the same test multiple times, this may lead to retest effects, namely an increase in performance due to participants' increased familiarity with the instrument (Scharfen, Peters, & Holling, 2018). An alternative would be to design different test versions to reduce any testing effects. Test versions could be designed to be of similar or increasing difficulty, as to account for students' expected increase in language ability over time. This approach requires test design procedures that allow the researchers to statistically equate different test versions, so that scores from different test forms measuring the same skill can be converted into a single scale (Goldstein, 1982). In this regard, the Rasch model represents a suitable methodological and analytical framework to design, validate, equate and analyse research instruments for longitudinal research (Kasprowicz et al., forthcoming).

1.4. Principles of Rasch analysis

Rasch measurements are "a family of probabilistic models that are used to predict the outcome of encounters between persons and assessment/survey items" (Aryadoust, Ng, & Sayama, 2021, p. 7). Rasch models have become increasingly adopted not only in language assessment research (Aryadoust et al., 2021; Dunn, 2024), but also for survey design and validation (see, for example, Leeming & Harris, 2024; Phipps, 2023; Yamashita, 2022 in this journal). A key principle of the Rasch model is that the items in a test are not assumed to be of equal difficulty (Boone & Noltemeyer, 2017). Therefore, Rasch analysis estimates both the difficulty of individual items and the ability of respondents onto the same scale. Since the Rasch model assumes that test performance is the result of a latent variable that the test aims to measure (Aryadoust et al., 2021), person abilities represent a measure of how much of the latent trait each respondent possesses.

The Rasch model and its analysis can be used for test validation as well as test equating. Various approaches for equating tests exist, all based on the principle that test versions need to share some information in the form of items or test-takers. In this study, we consider the common item equating procedure, whereby two or more tests share a number of common items spread along the difficulty continuum (Linacre, 2023). Through Rasch analysis, it is possible to both equate and validate two or more test versions sharing common items, so as to generate a common scale on which item and person measures from both tests are estimated.

1.5. Study aims

This research is part of the Progression in Primary Languages (PiPL) study, a longitudinal project tracking primary school children's progression in French, German and Spanish in England, as well as investigating how individual and contextual factors influence language learning. Using the design and validation process of the receptive vocabulary size tests as a case exemplar, this article intends to provide new perspectives on creating tests of LOTE for YLLs learning in input-poor, instructed contexts. After identifying potential population- and context-related threats to test reliability and validity using data from the early version of the tests, we showcase how these issues have been addressed and their impact on the psychometric properties of the revised instruments. The revised tests and the anonymised datasets are available on University of Reading's Research Data Archive (see Morea et al., 2024a). The revised tests are also available on the IRIS (Marsden & Mackey, 2016) repository (see Morea et al., 2024b).

2. The initial test

2.1. Context

In England, the context of this study, primary schools are required to teach a modern or ancient foreign language during the last four years of primary education, namely from Year 3 (students aged 7–8) to Year 6 (students aged 10–11). Schools choose the language to be taught, which, among modern foreign languages, tends to be French, German or Spanish (Collen & Duff, 2024). Whilst the National Curriculum specifies that students should "make substantial progress in one language" (DfE, 2013, p. 2), it does not provide any guidance regarding what specific content and structures should be taught. Additionally, no national assessment procedures exist to evaluate language learning, and school-based approaches to tracking students' linguistic progression are variable and ad-hoc, partly due to the lack of primary school teachers who specialise in L2 teaching (McLachlan, 2009).

Given this lack of curriculum guidance, schools are left with the task of deciding what structures to teach and when these should be introduced, and thus which Scheme of Work (SoW) to adopt. A SoW is a plan detailing the sequence of teaching and learning activities, as well as the amount of time devoted to each topic and the procedures used to evaluate whether the learning objectives have been achieved (Wallace, 2014). Whilst some schools may design their own SoW, SoWs for French, German and Spanish are also commercially available (free and paid-for), some of which are widely adopted across the country (Collen & Duff, 2024; Kasprowicz & Graham, in progress).

The picture is further complicated by variability in the amount of time schools allocate to language learning (Collen, 2022) and in teacher expertise in the language being taught (McLachlan, 2009). As a result, linguistic progression and vocabulary growth are likely to be very slow compared to other contexts where students are also exposed to the language of instruction outside the school environment (as is the case with English) (Graham et al., 2017). The resulting variability in student progression has implications for the transition to secondary school, as secondary language teachers may seek to address this by ``simply reteaching what was meant to have been covered in earlier years'' (Graham et al., 2017, p. 924).

To reduce variability in language provision within our study, we recruited schools who allocate 45–60 min to language learning per week and which follow a systematic and defined SoW. However, participating schools still differed in terms of which specific SoW they used. Further, there was variation in teacher expertise across participating schools, as some had a specialist language teacher responsible for the teaching of French, German or Spanish, whilst in other schools the individual classroom teacher was responsible for L2 teaching, regardless of whether they had ever learnt the language being taught.

2.2. Initial test design

2.2.1. Format

The initial tests consisted of three test versions for each language (French, German, Spanish) each containing 25 multiple-choice items, of which five items were common across test versions. The format of the tests was inspired by Nation and Anthony's (2016) Picture Vocabulary Size Test. Each item consisted of a word in the target language (a noun, verb or adjective), which students read and heard twice, followed by four options in English, represented in both written and pictorial form (Fig. 1). Each test began with a simple example to ensure children understood the task. The tests were designed to take 10–15 min to complete and to be administered twice per year, once at the beginning of the second school term (January–February) and once towards the end of the school year (June–July), using a different test version at each timepoint.

2.2.2. Content

Due to the absence of corpora drawing on natural classroom interactions in French, German and Spanish in low-input English primary school settings, as well as the lack of existing instruments designed for YLLs in England, a two-pronged approach was used for creating a reference wordlist from which the test items were sampled.

In the first instance, the Routledge frequency-based wordlists were consulted to identify the top 2000 most frequent lemmas in French (Lonsdale & Le Bras, 2009), German (Tschirner & Möhring, 2019) and Spanish (Davies & Davies, 2018). However, to limit the potential bias introduced by frequency-based vocabulary selection (see Section 1.2), we also identified low-frequency French, German and Spanish words commonly taught in primary school contexts in England.

As previously discussed, the English national context is characterised by heterogeneity in curriculum content due to the lack of a national programme of study detailing the language to be taught in primary schools. However, both free and paid-for schemes of work (SoW) exist in all three languages. Drawing on the findings of an online questionnaire distributed to primary language teachers in England in 2021 (Kasprowicz & Graham, in progress), the four SoWs most commonly used for language teaching by primary school teachers ($n = 151$) in the survey were identified for each language.[1] A vocabulary list was created that included all target language lemmas in these SoWs, together with information on the number of SoWs in which each lemma appeared. Table 1 reports a breakdown of the number of lemmas by number of SoWs and frequency band. As shown in Table 1, a noticeable proportion of the lemmas appearing in the SoWs were in fact lower-frequency words (>2000), thus supporting our two-pronged approach to vocabulary selection.

[1] SoWs used: Early Start (German), Goethe Institute (German), Language Angels (French, Spanish), Lightbulb Languages (French, Spanish), Primary Languages Network (French, German, Spanish), and Rachel Hawkes' KS2 Languages (French, German, Spanish).

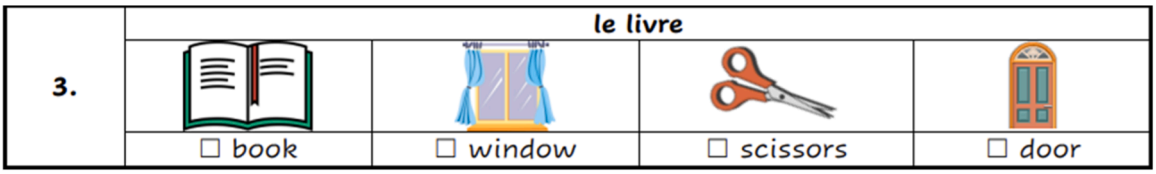

Fig. 1. Example of a vocabulary test item (from revised French test).

Table 1
Words included in commonly used SoWs (nouns, adjectives and verbs only), by frequency.

	French					German					Spanish				
	4/4	3/4	2/4	1/4	Total	4/4	3/4	2/4	1/4	Total	4/4	3/4	2/4	1/4	Total
<2000	68	61	84	177	390	89	68	112	198	467	67	60	102	216	445
>2000	21	49	99	327	496	15	36	96	0	147	22	39	84	0	145
Total	89	110	183	504	886	104	104	208	198	614	89	99	186	216	590

Note. Row 2: number of SoWs in which a lemma appeared. Column 1: frequency band.

It is also important to note that not all words appearing in the top 2000 frequency band are words that beginner, primary-school aged language learners would be expected to have encountered in the classroom (e.g., French: juger—to judge, frequency: 395; German: das Unternehmen—firm, company, frequency: 291; Spanish: fuerza—strength, frequency: 290). Therefore, the top 2000 words for each language were systematically filtered to identify those relevant to primary school language learning. Specifically, a top 2000 word was excluded from our reference list if it did not appear in either:

- One or more of the four commonly used SoWs identified for that language, or
- In the French/German/Spanish vocabulary lists of the Foundation tier General Certificate of Secondary Education (GCSE), the first academic qualification that students can obtain in a language during secondary education in England (typically age 15–16).

The final wordlists comprised the filtered top 2000 lemmas plus the lower-frequency lemmas appearing in at least two of the four SoWs for each language (Table 2). From these lists, test items were randomly selected using stratified sampling, so as to maintain the same proportion of lemmas by word class (nouns, adjectives and verbs) and frequency band, as in the reference lists and across test versions.

A final consideration when sampling vocabulary was the treatment of cognate words, namely words from one language that have a form-similar translation in a different language (Schepens, Dijkstra, & Grootjen, 2012). The presence of cognates within vocabulary size tests can be problematic. On the one hand, items containing cognates are easier to answer (Jordan, 2012); on the other hand, the ability to correctly recognise the meaning of a cognate may not solely be an effect of L2 receptive vocabulary size, but also of L1 knowledge (Allen & Nakamura, 2023), thus potentially threatening construct validity. In this study, we systematically identified cognates based on the normalised Levenshtein distance, a coefficient indicating the degree of orthographic similarity between pairs of words from two languages, adjusted for word length using the equation proposed by Schepens et al. (2012), as well as their proposed coefficient of 0.49 as the upper limit to consider a word pair as cognates. In each test version, we included three cognates (one per word class, 12 % of the test items), so that children with limited receptive L2 vocabulary knowledge were likely to recognise at least some of the vocabulary in the tests, thus resulting in a more motivating test experience.

2.3. Test administration and ethics

The study received ethical approval from the University of Reading. The initial tests were administered in February and March 2023 in 18 English primary schools (French: $n = 7$, participants: $n = 640$; German: $n = 3$, participants: $n = 291$; Spanish: $n = 8$, participants: $n = 731$) to children in Year 3 (7–8 years old), Year 4 (8–9 years old) and Year 5 (9–10 years old). Year 6 students (aged 10–11, final year of primary school) were not included in the first year of data collection as no longitudinal data for these students could be obtained in successive years. The lower number of schools teaching German may be explained by the overall national decline in schools teaching the language at primary and secondary level.

All schools volunteered to participate in the PiPL project with informed consent provided by the head teacher and class teachers. An information sheet was distributed to the families of all students in the target year groups, detailing the aims of the project, its timeline and the research activities involved. A simplified information sheet and assent form was also provided to and completed by each student-participant. A short video, introducing the project team and the aims of the project was shared with each school and shown to participants, prior to completing the assent form and research activities.

The research team visited each participating school to administer the language tests in the classroom during regular teaching hours using a pen-and-paper format. At the beginning of the session, the researcher reminded participants of the project aims, with the tests presented as language activities that the researchers would use to understand what the children had been learning in the target

Table 2
Number of selected lemmas by frequency band and language (initial test).

	French	German	Spanish
Top 2000	798	851	1004
>2000	169	147	145
Total	967	998	1149

language. It was emphasised that students should not worry if they did not know some of the language featured in the activities, and that neither their teacher nor their parents would be able to access their individual responses. The researchers used a Power Point presentation to guide students through the tests. After test administration, each participant was assigned a unique code, and all identifying information was removed in the data cleaning process.

2.4. Scoring and data analysis

Table 3 reports the number of participants who completed the vocabulary tests in the first round of data collection. Student responses were initially entered into Excel files. Although the multiple-choice format of the test made the data entry process straightforward, occasional instances of ambiguous responses were flagged and discussed within the team until a final decision was reached. The Excel files were imported into the software SPSS (version 29). Responses were systematically scored into binary data (correct = 1, incorrect = 0) using the "convert into a different variable" function. Multiple responses and non-systematic missing answers were coded as incorrect. The SPSS datasets were then imported into the programme Winsteps (version 5.7.2) to evaluate the psychometric properties of the tests using Rasch analysis.

2.5. Results

2.5.1. Descriptive statistics

Table 4 provides an overview of the distribution of the total scores of each test version, whereas Table 5 displays the average test scores and percentage of correct answers by year group and language. The distribution of total scores was normal as Skewness and Kurtosis values fell within Hair et al.'s (2010) recommended range (Skewness: −2 to +2, Kurtosis: −7 to +7). The first noticeable characteristic of the data was that the mean percentage of total correct answers across all tests was just over 41 % (min: 36.52 %, max: 49.09 %). This suggests that the tests were noticeably difficult for participants, especially considering the relatively high probability of correctly answering any given item by chance (25 %). Additionally, differences in average scores by students in their first, second and third year of language learning were small, either indicating that linguistic progression was limited or that the tests were not sensitive enough to capture meaningful differences in receptive vocabulary knowledge across year groups.

2.5.2. Rasch analysis: person and item reliability

As previously mentioned, Rasch analysis estimates the measures of both persons and items, as well as providing a reliability and a separation coefficient for each. Person measures refer to the estimated ability level of the test takers, whereas item measures refer to the difficulty of the item. Person and item separation indicate the number of "statistically different levels of item difficulty or person ability in the data" (Aryadoust et al., 2021, p. 11), and they can thus be used to evaluate the spread in test-taker performance and item difficulty. Finally, the reliability coefficient of both persons and items can be interpreted similarly to a Cronbach's alpha coefficient. The reliability indices represent the likelihood that high and low ability measures were indeed estimated for high and low ability test takers, respectively (Aryadoust et al., 2021). However, low person reliability indices may either reflect little variation in the data due to homogeneity of ability levels within the sample (Aryadoust et al., 2021; Linacre, 2023) or be due to the number of items being insufficient to discriminate between different ability levels within the sample (Linacre, 2023).

Table 6 reports person and item reliability and separation coefficients, together with the average person and item measures for each test. Whilst the item reliability was consistently excellent across test versions (0.88–0.99), thus indicating both reproducibility of the item measures and a good range of statistically significant item difficulty levels in the test (Aryadoust et al., 2021), person reliability was concerningly poor, with coefficients as low as 0.16.

The mismatch between person and item measures was confirmed after inspection of a Wright map for each test version. A Wright map is a visual representation of the location of items and persons on the same ability/difficulty logit scale. In a well-balanced test, one would expect a similar distribution between item difficulties and person abilities. However, the Wright maps of the initial tests

Table 3
Participants by language learnt and class (initial test).

	Number of year 3 students	Number of year 4 students	Number of year 5 students	Total number of students
French	228	214	198	640
German	97	94	100	291
Spanish	247	218	266	731
Total	572	526	564	1662

Table 4
Descriptive statistics of total test scores (0–25) by test version (initial test).

French						
t. v.	N (total)	n (valid)	M	SD	Skewness	Kurtosis
A	153	138	10.19	2.73	−0.17	−0.66
B	194	160	9.96	2.85	0.73	2.87
C	293	269	10.55	2.89	0.68	2.68
German						
	N (total)	n (valid)	M	SD	Skewness	Kurtosis
A	65	50	9.64	2.97	−0.14	−0.85
B	79	75	12.27	3.12	−0.45	−0.19
C	147	122	10.89	2.37	0.57	−0.27
Spanish						
	N (total)	n (valid)	M	SD	Skewness	Kurtosis
A	162	151	9.38	3.09	0.22	−0.31
B	189	182	9.13	2.58	0.41	1.22
C	380	359	10.70	3.03	0.93	3.38

Note. "t.v." in column one indicates the test version. Differences between "total N" and "valid n" are due to some participating students either being absent on the day of data collection or joining/leaving the session whilst the test was being administered.

Table 5
Average test scores by language and year of language study (initial test).

	French			German			Spanish		
	n (valid)	M	%	n (valid)	M	%	n (valid)	M	%
Whole sample	567	10.29	41.16	247	11.06	44.24	692	10.00	40.00
1st Year	209	9.52	38.08	80	10.28	41.12	254	9.25	37
2nd Year	193	10.54	42.16	74	11.18	44.72	192	10.01	40.04
3rd Year	165	10.99	43.96	93	11.63	46.52	246	10.78	43.12

Table 6
Person and item reliability and measures (initial test).

French				
	Person measures		Item measures	
Test version	Reliability	Separation	Reliability	Separation
A	0.31	0.67	0.96	4.67
B	0.43	0.86	0.96	4.83
C	0.47	0.95	0.98	7.98
German				
	Person measures		Item measures	
Test version	Reliability	Separation	Reliability	Separation
A	0.40	0.81	0.88	2.65
B	0.50	1.01	0.94	4.01
C	0.16	0.44	0.97	5.91
Spanish				
	Person measures		Item measures	
Test version	Reliability	Separation	Reliability	Separation
A	0.51	1.02	0.97	6.01
B	0.28	0.63	0.96	4.83
C	0.53	1.07	0.99	8.32

indicated that the test items were considerably more difficult than the participants' estimated ability levels, likely resulting in a large amount of error in the data due to guessing. As an example, Fig. 2 shows the Wright map of version A of the French tests. The distribution of the items by difficulty can be seen on the right-hand side of the scale (with more difficult items appearing towards the top of the scale), and the distribution of participants by their ability level are shown on the left-hand side of the same scale. The map clearly shows that most of the test items were targeted to higher ability levels than those present in the participant sample, except for two

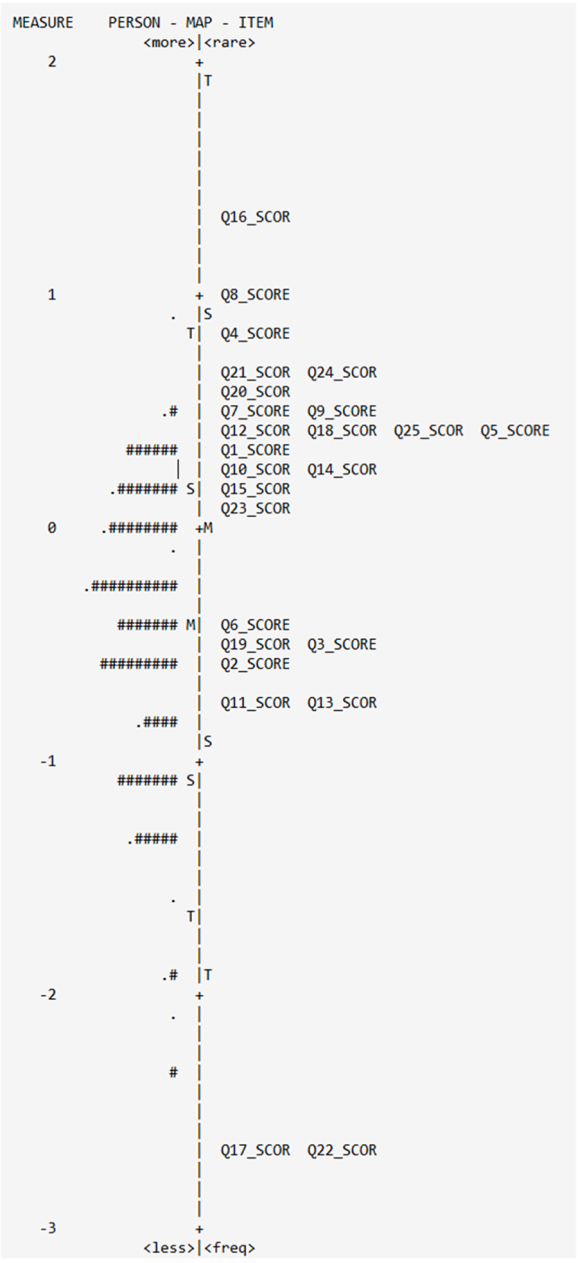

Fig. 2. An example wright map (version a of the initial French vocabulary test).

extremely easy items located towards the bottom of the scale (items Q17 and Q22), which were both cognate words.

In consideration of the poor person reliability of the initial instruments, it can be assumed that test validity was also not acceptable. In the next section, the possible reasons behind the poor psychometric properties of the initial instruments are discussed, together with the actions that were taken to address these issues.

2.6. Test revisions: issues, approach, revised test content and format

Based on the results of the descriptive statistics and Rasch analysis, three possible issues were identified as potential causes of the poor person reliability:

1. The initial tests were too difficult for the sample and target population, resulting in a large amount of noise in the data due to guessing. This was clear from both the overall test scores and the item difficulty and person ability distribution on the Wright maps.

2. The sample and, by extension, the population of reference may truly be homogeneous in terms of receptive vocabulary skills. If this were the case, it could be an indication of limited progress in receptive vocabulary knowledge during primary education.
3. The number of test items was insufficient. To be able to track relatively small variations in ability levels in a homogeneous population, a larger number of items may be needed.

Each of these issues was tackled during the second phase of test design, and the actions taken are summarised in Table 7.

Firstly, test difficulty was likely driven by the inclusion of highly frequent lemmas (top 2000 frequency band) that nonetheless appeared in none or only a few of the reference SoWs. We therefore hypothesised that these lemmas may have been unfamiliar to most participants and therefore of a high difficulty level. In response, new reference wordlists were created for each language following a SoW-based approach to vocabulary selection. Accordingly, the revised wordlists only included words common across at least three of the four identified SoWs for each language, irrespective of word frequency. As a result, the revised reference wordlists contained a smaller number of lemmas, which, however, were more representative of the typical language children are exposed to in the primary school languages classroom (French: $n = 199$, German: $n = 207$, Spanish: $n = 188$). From the revised reference wordlists, items were sampled using a similar approach to the initial tests, namely through stratified random sampling after controlling for word class, word frequency and number of SoWs in which each item appears (3/4 or 4/4). This was to ensure that the test content reflected the characteristics of the reference wordlists. Finally, the revised tests included the same number and type of cognates as the initial tests.

The issue of homogeneity in sample ability was tackled by including students in the last year of primary education (Year 6, children aged 10–11). Since the revised tests were administered one year after the initial tests, the Year 6 children were students who were in Year 5 when the initial tests were administered. As a result, a new phase of participant recruitment was conducted to include the new Year 3 students in each partner school. Additionally, the number of test versions per language was reduced from three to two, resulting in a larger sample of children completing each version.

Finally, the number of items in each test was expanded from 25 to 40, thus considerably increasing test length. Since, as discussed in the literature review, long tests are not appropriate for YLLs, the format of test administration within the PiPL project was re-designed. Firstly, the number of times individual tests would be administered per academic year was reduced from two to one, resulting in additional time available to administer each test. Secondly, and most importantly, the revised vocabulary tests were separated into two parts of 20 items each. The first part was presented at the beginning of the 60-minute data collection session, whilst the second part was administered during the second half of the session (immediately after a short movement break during which children engaged in a game with the researcher and classroom teacher). This approach was used to limit participant fatigue and sustain motivation during the test.

3. The revised test

3.1. Test administration and data analysis

The revised tests were administered in January and February 2024 in 17 primary schools in England to children from Year 3 to Year 6 (7–11 years old). Table 8 reports the sample size by test language and year group. New participants (i.e., Year 3 students and any children who joined the schools after the initial test administration) were recruited following the same procedures explained in Sections 2.2 and 2.3. The approach to test administration, data entry, coding and analysis was analogous to the previous round of data collection.

3.2. Results

3.2.1. Descriptive statistics

Table 9 presents the distribution of total scores on the two test versions for each language, and Table 10 presents the average score and percentages by language and year of language learning. The distribution of total scores was normal as Skewness and Kurtosis values fell within the recommended range of ± 2 and ± 7, respectively. The percentage of total correct responses suggests that the revised tests may have been easier than the initial tests, although the higher mean percentages may also be driven by the inclusion of older students (the Year 6 classes). Compared to the initial tests (see Table 5), the revised tests seemed to better capture progression in test performance among older students. For example, students in their fourth year of French learning were able to recognise an average of nearly eight more words compared to students in their first year, an increase of 19.6 % in test performance. However, these differences were still arguably small, thus reinforcing the hypothesis that progression in vocabulary knowledge may be slow.

3.2.2. Rasch analysis

Following the guidance by Aryadoust et al. (2021) on rigorous reporting of Rasch analysis, the next subsections present the results of person and item fit, person and item reliability, and construct validity (i.e., item dimensionality).

3.2.2.1. Item and person fit. In Rasch analysis, infit and outfit statistics are used to evaluate the goodness of fit of individual items and persons (i.e., the extent to which item and response patterns align with the Rasch model). Both statistics can reveal anomalies in item performance or person response patterns. The infit and outfit statistics usually reported are the mean square index and the standardized *Zstd* metrics, although the latter is recommended when $N < 250$ (Aryadoust et al., 2021). Mean square indices have an

Table 7
Overview of initial test issues and actions taken.

Aspect	Initial tests	Issue	Action	Revised tests
Vocabulary identification	Vocabulary for reference wordlists selected from: 1. Top 2000 most frequent lemmas, filtered to lemmas that appeared in: • at least 1/4 SoWs, and/or • the GCSE Foundation vocabulary lists. 2. List of lower-frequency lemmas (>2000) appearing in at least 2/4 SoWs.	Tests too difficult for target sample/population.	Changed the approach for vocabulary identification to more closely reflect the language children are likely to be exposed to in primary language teaching.	Reference wordlists reduced to include only lemmas appearing in 3/4 or 4/4 SoWs, regardless of word frequency.
Sample	Year 3 to Year 5 (children aged 7–10)	Population ability may be homogeneous.	Extended the sample to students in Year 6 (children aged 10–11). Reduced number of test versions from 3 to 2 to increase sample size.	Year 3 to Year 6 (children aged 7–11). Two test versions per language.
Test items	25 items/test (5 common items). Two test administrations per year.	Insufficient number of test items.	Increased the number of items and reduced the number of administrations to once per year.	40 items/test (10 common items).

Table 8
Participants by language learnt and class (revised test).

	Number of year 3 students	Number of year 4 students	Number of year 5 students	Number of year 6 students	Total number of students
French	168	243	228	205	844
German	85	132	129	134	480
Spanish	160	229	238	251	878
Total	413	604	595	590	2202

Table 9
Descriptive statistics of total test scores (0–40) by test version (revised test).

French							
Test version	N (total)	N (valid)	M	M%	SD	Skewness	Kurtosis
A	442	381	21.59	53.98	5.93	0.19	−0.32
B	402	361	22.26	55.65	6.15	−0.11	−0.39
German							
	N (total)	N (valid)	M	M%	SD	Skewness	Kurtosis
A	233	196	21.89	54.73	5.89	0.13	0.35
B	247	217	22.78	56.95	4.59	−0.34	−0.21
Spanish							
	N (total)	N (valid)	M	M%	SD	Skewness	Kurtosis
A	419	367	17.21	43.03	5.19	0.20	0.18
B	459	412	18.47	46.18	5.67	0.42	0.71

Table 10
Average test scores by language and year of language study (revised test).

	French			German			Spanish		
	n (valid)	M	%	n (valid)	M	%	n (valid)	M	%
Whole sample	742	21.92	54.80	413	22.36	55.90	779	17.88	44.70
1st Year	148	17.18	42.95	82	17.54	43.85	146	15.84	39.60
2nd Year	236	21.66	54.15	103	21.36	53.40	211	16.89	42.23
3rd Year	189	23.18	57.95	108	23.87	59.68	207	19.29	49.23
4th Year	169	25.02	62.55	120	24.29	60.73	215	18.87	47.18

expected value of 1.0 (Aryadoust et al., 2021), with higher values indicating that the person or item underfit the model and lower values indicating that the person or item are overfitting (i.e., their pattern are too predictable). High mean square indices are considered a greater threat to measurement accuracy than low mean square indices (Aryadoust et al., 2021; Linacre, 2002). Various cut-off criteria to identify misfitting items and persons are recommended in the Rasch literature; here, we applied the sample-size-dependent formula recommended by Aryadoust et al. (2021) to determine the limit of acceptable mean square values.

In each test, several items were identified as misfitting (Table 11) and were removed from further analysis. Interestingly, most of these misfitting target words were common across only three of the four SoWs from which words were sampled. As a result, the misfitting patterns of participant responses to these items may be explained by the fact that students in some participating schools might have never encountered these words before.

The same criteria used to identify misfitting items were also adopted to investigate misfitting persons in each dataset. To decide whether misfitting persons should be systematically excluded, Linacre (2010) recommends cross-plotting the person ability measures of the dataset including the misfitting persons with the same measures from a dataset excluding the misfitting persons. If the slope of the best-fit trend line closely approximates 1 (i.e., the slope of the identity line), then the misfitting persons may be retained in the analysis, since their inclusion does not affect the estimation of person ability measures. In all cases, the slope of the best-fit trend lines was virtually identical to the slope of the identity lines (range of slopes: 0.999726–0.999989). This gave us confidence that the presence of the misfitting persons in the datasets did not degrade person measure estimates, and the misfitting persons were thus not removed from the datasets.

3.2.2.2. Person and item reliability. The person and item reliability statistics of the revised tests were calculated for each test version, as well as for the combined tests after common item equating (Table 12). Firstly, a comparison between Table 5 (initial tests) and Table 9 (revised tests) reveals that the average person measures are consistently higher than the initial tests, indicating that the revised tests were overall easier compared to the initial tests. Furthermore, both person and item reliability have improved. As a rule of thumb, a person reliability coefficient of 0.80 or above is regarded as good, although values between 0.70 and 0.79 may be considered

Table 11
Misfitting items (revised test).

Test	Item	Infit cut-off	Infit MnSq	Outfit cut-off	Outfit MnSq	Target word	Answer	SoWs, N
French A	20	1.11	1.29*	1.31	1.39*	la terre	earth	3
	10		1.23*		1.35*	la natation	swimming	4
	18		1.17*		1.25	l'épaule	shoulder	3
	33		1.12*		1.14	il neige	it is snowing	3
French B	27	1.11	1.16*	1.31	1.61*	le petit-déjeuner	breakfast	3
	40		1.22*		1.44*	lent	slow	3
	34		1.12*		1.29	je joue	I am playing	4
	14		1.16*		1.18	le gâteau	cake	3
German A	23	1.14	0.98	1.43	2.19*	das Tier	animal	3
	40		1.2*		1.54*	langsam	slow	3
	20		1.29*		1.38	das Frühstück	breakfast	4
German B	33	1.13	1.22*	1.40	1.37	ich wiederhole	I repeat	3
	28		1.15*		1.23	das Auge	eye	3
	14		1.14*		1.33	der Fluss	river	3
Spanish A	12	1.10	1.05	1.31	1.63*	la habitación	room	3
	14		1.15*		1.36*	la natación	swimming	4
	20		1.15*		1.31	las gafas	glasses	3
	39		1.11*		1.15	alto	tall	4
Spanish B	23	1.10	0.99	1.30	1.56*	el coche	car	3
	39		1.16*		1.36*	viejo	old	3
	13		1.16*		1.28	medianoche	midnight	3
	15		1.15*		1.21	el pelo	hair	3
	12		1.11*		1.13	el colegio	school	3

* *Note*. indicates mean square (MnSq) values exceeding the cut-off value.

Table 12
Person and item reliability and measures (revised test).

French				
	Person measures		Item measures	
Test version	Reliability	Separation	Reliability	Separation
A&B	0.80	2.00	0.99	8.60
A	0.80	1.99	0.99	8.68
B	0.80	2.01	0.99	8.13
German				
	Person measures		Item measures	
Test version	Reliability	Separation	Reliability	Separation
A&B	0.75	1.72	0.98	7.28
A	0.79	1.94	0.97	6.12
B	0.69	1.50	0.98	7.72
Spanish				
	Person measures		Item measures	
Test version	Reliability	Separation	Reliability	Separation
A&B	0.75	1.75	0.98	7.80
A	0.72	1.61	0.98	7.51
B	0.79	1.84	0.98	7.70

Note. A&B indicate the combined tests.

acceptable (Aryadoust, 2013; Bond, Yan, & Heene, 2020; Xing, Liu, Li, Cui, & Biering-Sørensen, 2024). Based on this, the person reliability coefficients of the revised tests lie within an acceptable range (average: 0.74; minimum: 0.69; maximum: 0.80). This is also evidenced by an increase in person separation indices, which in some cases approaches 2. However, whilst the French tests could separate the sample into two statistically distinct ability levels, the German and Spanish tests could not.

On the one hand, the revised approach to vocabulary sampling was effective in improving the reliability of the tests and in decreasing their difficulty. By reducing the pool of lemmas as to only include words recurring across the most commonly used schemes of work, the revised tests seemed to more closely reflect the language students were exposed to in the classroom. Nonetheless, the modest reliability coefficients and the limited spread of ability levels provide clear indication of a homogeneous sample, suggesting that vocabulary learning in this particular context may be slow and limited. In consideration of the constraints of measuring linguistic progression in input-poor contexts characterised by limited language teaching and variability in teacher expertise, the tests can be considered sufficiently reliable for this context.

3.2.2.4. Construct validity and dimensionality. Construct validity refers to the extent to which a research instrument measures the construct or domain it purports to measure (Cohen et al., 2007). In Rasch analysis, construct validity may be assessed by inspecting the dimensionality of a test. If the underlying construct is theorised to be unidimensional (as it is the case in this study), the degree of unidimensionality of the instrument may be assessed via principal component analysis of residuals (PCAR) (Aryadoust et al., 2021; Linacre, 2023). This is also one of the assumptions of the Rasch model. Through PCAR, researchers can evaluate the presence of substantive additional dimensions in the item residuals beyond the main Rasch dimension, which in turn would point to violations of the unidimensionality assumption. Linacre (2023) argues that no instrument is perfectly unidimensional, and that researchers should instead assess whether the lack of unidimensionality is "sufficiently large to threaten the validity of […] results" (p. 638). In this regard, Linacre (2023) recommends several approaches to assess the severity of undimensionality violations:

- Assessing the contrast eigenvalue: if a secondary dimension (or, in Rasch analysis terms, a *contrast*) has an eigenvalue >2, this dimension may be substantive and warrant further investigation (Aryadoust et al., 2021; Linacre, 2023).
- Assessing the size of the secondary dimension relative to the Rasch dimension: if the variance explained by the secondary dimension is considerably smaller than the variance explained by the Rasch dimension, the effect of the secondary dimension on the instrument may be negligible. However, no criteria exist on how much larger the Rasch dimension should be compared to a contrast to reject multidimensionality.
- Inspecting the disattenuated correlations of person measures: if the correlation between person measures on a cluster of items loading on the secondary dimension and another cluster is weak or negative, then the secondary dimension may reflect something different than the Rasch dimension.
- Examining the items with strong loadings on a secondary dimension may help to decide whether a contrast represents a distinct dimension or a *strand* of the Rasch dimensions (e.g., a cluster of items testing subtraction within an arithmetic test). In the latter case, the unidimensionality assumption may still hold.

To check the degree of multidimensionality, PCAR was conducted on each test version. Firstly, the raw variance explained by measures (i.e., the amount of variance explained by the Rasch dimension) was consistently larger than the minimum recommended value of 20 %, (range: 22.4 %−34.6 %) (Reckase, 1979; Wind & Hua, 2022).

Table 13 reports the characteristics of the first two contrasts (i.e., potential sub-dimensions) in each test language and version, as well as the decision to ignore or further investigate each contrast. Across all tests, all contrasts beyond the second were negligible (eigenvalue <2). The first contrast of the French version B test, as well as the first contrast of the Spanish version B test, were further examined as potential threats to the unidimensionality assumption and construct validity. These contrasts were deemed to warrant further investigation in consideration of their eigenvalues (2.77 and 2.61, respectively), the relatively smaller ratio between the Rasch and contrast dimensions (5.89 and 3.90, respectively) and a weak disattenuated person correlation between item clusters (0.27 in both cases).

Table 13
Overview of secondary contrasts and characteristics, by language and test version (revised test).

French												
	Version A						Version B					
	Eig.	Ratio	Disatt. Pers. Corr.			Flag?	Eig.	Ratio	Disatt. Pers. Corr.			Flag?
			1–2	1–3	2–3				1–2	1–3	2–3	
1st c.	2.35	5.89	0.85	0.40	0.78	N	2.77	4.63	0.68	0.27	1	Y
2nd c.	1.95	7.10	0.83	0.59	0.99	N	1.98	6.47	1	0.75	0.91	N
German												
	Version A						Version B					
	Eig.	Ratio	Disatt. Pers. Corr.			Flag?	Eig.	Ratio	Disatt. Pers. Corr..			Flag?
			1–2	1–3	2–3				1–2	1–3	2–3	
1st c.	2.07	6.57	0.91	0.39	0.80	N	2.50	7.81	1	1	0.43	N
2nd c.	1.96	6.94	0.80	0.40	0.92	N	1.93	10.12	0.83	0.28	0.81	N
Spanish												
	Version A						Version B					
	Eig.	Ratio	Disatt. Pers. Corr.			Flag?	Eig.	Ratio	Disatt. Pers. Corr.			Flag?
			1–2	1–3	2–3				1–2	1–3	2–3	
1st c.	1.83	5.67	0.66	0.56	0.99	N	2.61	3.90	0.90	0.27	0.80	Y
2nd c.	1.63	6.36	0.80	0.59	0.89	N	1.83	5.54	1	0.62	0.80	N

Note. Column 1 = contrast. Colum 2 = contrast eigenvalue. Column 3 = eigenvalue ratio between Rasch dimension and contrast (e.g., 5.89 = the Rasch dimension is 5.89 times the size of the contrast). Column 4 = disattenuated person correlation between the three item clusters within each contrast. Column 5 = whether a contrast warranted further investigation.

For each contrast, the dimension-defining items were examined to understand the nature of the contrast. In all cases, almost all of the items most strongly and positively loading on the three contrasts consisted of lemmas appearing in three out of four SoWs, whilst the items most strongly and negatively loading on the contrast consisted of lemmas appearing in all four SoWs (Table 14). Based on this, it can be argued that the secondary dimensions reflected the inclusion of vocabulary that may not have been introduced in some of our participating schools. As a result, students from schools in which these lemmas had been introduced were at an advantage over other participants, and the resulting response patterns formed a secondary dimension in some of the datasets. Since these secondary dimensions reflect the way in which items were sampled rather than a distinct ability trait that those items were testing, it can be concluded that the unidimensionality assumption was not violated and that all tests appeared to measure a single latent trait.

Finally, no issue of local independence was found in any tests, which further confirms that the unidimensionality assumption was not violated. Local independence refers to the Rasch assumption that "after conditioning for the latent trait, performance on one test item does not covary with performance on other items" (Aryadoust et al., 2021, pp. 8–9). This assumption can be checked by examining the correlation matrix of the item residuals. Concerns about violation of this assumption may arise if the residuals of any pair of items exhibit a strong correlation, usually above 0.50 (Aryadoust, 2013) or 0.70 (Linacre, 2023). None of the tests presented issues of local dependence as all correlation coefficients were small (largest $r = 0.32$).

4. Discussion and conclusions

The purpose of this article was to offer new perspectives on approaches to develop language assessment instruments for YLLs in input-poor, instructed contexts. To this end, we discussed the design and validation of a novel receptive vocabulary size test aimed at longitudinally tracking primary school children's vocabulary knowledge in French, German and Spanish in England. Two phases of instrument design, administration and validation were conducted to ensure the tests achieved satisfactory psychometric properties. The initial tests presented issues that resulted in poor person reliabilities, and most notably the inclusion of language that students were unlikely to have encountered in the classroom. This was reflected in misalignment between test difficulty and student ability resulting in a large amount of noise in the response patterns, likely due to guessing. Revisions to the way test language/items were selected, as well as changes in test format and administration considerably improved the instruments, as evidenced by their considerably improved person reliabilities and excellent item reliabilities, together with an overall close fit to the Rasch model.

The study results demonstrate that designing reliable and valid instruments to capture YLLs' linguistic progression in input-poor instructed contexts is possible but requires careful consideration of population- and context-related factors. The overarching challenge lies in the need to balance the statistical requirements of a reliable and valid test and the ethical needs of creating instruments accounting for YLLs' characteristics and closely reflecting their lived experience of the language being learnt (Hasselgren, 2012).

In the context of tests of receptive vocabulary size, test items are normally sampled from large banks of vocabulary (Laufer et al., 2004). One common approach is to sample items according to their frequency, based on the assumption that highly frequent words are more likely to be taught/encountered (Milton, 2008). However, this may not necessarily be the case, as the appropriateness of frequency lists will depend on the appropriateness of the corpora from which the words are extracted. Additionally, primary school teachers may prefer using topic-based approaches (whereby language content is introduced according to child-appropriate topics, rather than word frequency), which may result in the teaching of less frequent words Indeed, the present study's findings indicate that frequency may not represent a suitable criterion for vocabulary identification when testing beginner YLLs learning in an input-poor instructed context. This is evidenced by the clear improvement of person reliability in the tests after moving away from a frequency-based approach and instead prioritising recurrent vocabulary in commonly used schemes of work. The results suggest that an approach which accounts for the unique learning environment may be more suitable than a general frequency-based approach, particularly in contexts characterised by variability in curriculum content.

Test length represents another crucial aspect to balance when designing research instruments for YLLs. Whilst shorter tests are more suitable for this population of learners (Courtney & Graham, 2019; McKay, 2006), a large number of items may be needed to achieve satisfactory reliability and validity. This is particularly the case when using a multiple-choice design, as this item format is likely to introduce considerable noise in the data due to guessing. Moreover, an even larger number of items may be required when the sample and population of reference display homogeneous ability levels (Linacre, 2023), which seems to be the case with YLLs in England. The apparent paradox of conducting research on a diverse population with homogenous ability poses unique challenges for designing tests for YLLs. However, longer tests may still be ethically used with YLLs, as long as strategies are put in place to limit fatigue and sustain motivation.

Furthermore, Rasch analysis proved to be a suitable and useful methodological framework for designing, validating and analysing language tests for YLLs. Providing estimates of test-takers' ability as well as item difficulty allowed the researchers to evaluate the ability range within the sample and population of reference. Moreover, the possibility to equate different test versions sharing a nucleus of common items made Rasch analysis particularly useful for longitudinal and experimental research, where administering different versions of a test may be preferrable to reduce retest effects or to account for increases in language ability over time.

Finally, the proposed tests respond to the real need to develop language assessment instruments to monitor students' linguistic progress across primary schools in England (McLachlan, 2009). These tests have been designed and refined to tackle the contextual constraints specific to this national context, but which may also arise in other English-speaking contexts. These included: limited exposure to the target language, varying teacher expertise in the language, and variability in the linguistic structures and vocabulary being taught. As such, these tests are particularly suitable for administration across a range of primary schools in the country, regardless of teacher expertise and scheme of work used. In particular, they may be used to longitudinally track students' progression over the four years of language learning, thus providing valuable information to inform curriculum planning within individual primary

Table 14
Characteristics of items most strongly loading on the secondary dimensions (revised tests).

Test	Item number	Factor loading	Target word	Answer	Number of SoWs
French version B	3	0.59	le livre	book	3
	13	0.56	le cheval	horse	3
	12	0.50	le stylo	pen	3
	8	−0.56	le poisson	fish	4
	19	−0.45	la glace	ice cream	4
Spanish version B	31	0.59	Escucho	I listen	3
	2	0.56	el libro	book	3
	21	0.46	la casa	house	3
	28	0.46	el caballo	horse	4
	4	−0.45	Naranja	orange	4

schools. In this regard, different test versions should be administered alternately, in order to limit any retest effects that might result from readministering the same test multiple times. Through a core set of common items, pairs of test versions can then be equated onto a same measurement scale using Rasch analysis, making the two test forms comparable.

Additionally, these tests would enable researchers to investigate the nature and rate of student vocabulary learning in the three most commonly taught languages in the country. In turn, this would represent a first step not only to evaluate the suitability of the National Curriculum's objectives, but also to tackle the variability in student L2 ability when entering secondary education, which often results in secondary language teachers having to jeopardise any progression made during primary education by reteaching content that should have been taught in previous years (Graham et al., 2017). Administration of the tests in the final year of language learning (age 11), prior to children's transition to secondary school, would provide one source of vital information for secondary schools, to enable children's current level of language knowledge to be taken into account when planning language teaching in the early stages of secondary education. These tests, and particularly their design process, may also be adapted to other national contexts where LOTE are taught in primary school and where exposure to the language and amount of teaching time are limited.

The tests proposed in this study are not without limitations. Whilst the vocabulary included in the tests was likely to reflect the typical language taught in the primary languages classroom, there was no way to guarantee that the recurring lemmas from the four commonly used schemes of work were taught in each individual participating school. Specifically, the inclusion of lemmas appearing in most (i.e. three) but not all the identified common schemes of work meant that some vocabulary items may have been more difficult for some participants than others. This approach was however justified by the need to create a sufficiently large bank of vocabulary for meaningful assessment of children's receptive vocabulary knowledge within this context. Additionally, whilst some test forms displayed good reliability, this was not the case for all test versions. This can be explained by the difficulty in identifying language that was commonly taught across school contexts adopting different schemes of work. Nonetheless, in consideration of the constraints characterising the English context, we argue that the tests are sufficiently reliable and may thus be used for their intended purpose.

In conclusion, the study has shown the importance of accounting for population and contextual factors when designing language tests for young language learners. Moreover, it has produced a novel test of receptive vocabulary knowledge that may be employed nationally and globally to assess L2 knowledge among populations of learners that have traditionally received less research attention, and namely primary school children engaged in learning languages other than English. It is the authors' hope that this article will provide guidance on ways to tackle key challenges of longitudinally assessing YLLs' linguistic progression in input-poor instructed contexts. The next step will be to evaluate whether longitudinal test data confirms the preliminary evidence of homogeneity in primary school students' L2 receptive vocabulary knowledge, despite multiple years of mandatory language learning.

Funding sources

This work was funded through a UKRI Future Leaders Fellowship Grant, grant number: MR/V023470/1.

CRediT authorship contribution statement

Nicola Morea: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rowena Eloise Kasprowicz:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Astrid Morrison:** Writing – review & editing, Resources, Investigation, Data curation. **Carmen Silvestri:** Writing – review & editing, Resources, Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgements

We would like to thank Hannah Davidson, Heike Krüsemann, Jasmin Silver, as well as Alice Kirscher, Élise Freudenberg, Jiarun Ye, Noelia Gonzalez Hernandez and Xiaobo Li for their support with data collection and data entry. We would like to thank our partner schools, our participants and our contact teachers, as well as the authors of the Schemes of Work used in this study: Early Start, Goethe Institute, Language Angels, Lightbulb Languages, Primary Languages and Rachel Hawkes' KS2 Languages. We also thank the colleagues at the Institute of Education of the University of Reading for their helpful feedback on the article. Finally, we would like to thank Prof Aaron Batty for helping us with the preliminary Rasch analysis of the initial tests.

References

Allen, D., & Nakamura, K. (2023). The distribution of cognates and their impact on response accuracy in the EIKEN tests. *Language Testing, 40*(3), 771–795.

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Cambridge Scholars Publishing.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6–40.

Bailey, A. L., Heritage, M., & Butler, F. A. (2014). Developmental considerations and curricular contexts in the assessment of young language learners. In J. Kunnan (Ed.), *The companion to language assessment*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla079.

Bardel, C., Gudmundson, A., & Lindqvist, C. (2012). Aspects of lexical sophistication in advanced learners' oral production: Vocabulary acquisition and use in L2 French and Italian. *Studies in Second Language Acquisition, 34*(2), 269–290. https://doi.org/10.1017/S0272263112000058

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Taylor and Francis.

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Educational Psychology & Counselling, 4*. https://doi.org/10.1080/2331186X.2017.1416898

Butler, Y. G. (2019). Teaching vocabulary to young second- or foreign-language learners: What can we learn from the research? *Language Teaching for Young Learners, 1*(1), 4–33. https://doi.org/10.1075/ltyl.00003.but

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). Routledge Falmer.

Collen, I. (2022). *Language trends 2022: Language teaching in primary and secondary schools in England*. British Council. Retrieved from https://www.britishcouncil.org/sites/default/files/language_trends_report_2022.pdf.

Collen, I., & Duff, J. (2024). *Language trends England 2024: Language teaching in primary, secondary and independent schools in England*. British Council. Retrieved from https://www.britishcouncil.org/sites/default/files/language_trend_england_2024.pdf.

Courtney, L., & Graham, S. (2019). "It's like having a test but in a fun way": Young learners' perceptions of a digital game-based assessment of early language learning. *Language Teaching for Young Learners, 1*(2), 161–186. https://doi.org/10.1075/ltyl.18009.cou

Dabbagh, A., & Janebi Enayat, M. (2019). The role of vocabulary breadth and depth in predicting second language descriptive writing performance. *Language Learning Journal, 47*(5), 575–590. https://doi.org/10.1080/09571736.2017.1335765

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

Davies, M., & Davies, K. (2018). *A frequency dictionary of Spanish: Core vocabulary for learners* (2nd ed.). Routledge.

De Wilde, V. (2023). The auditory picture vocabulary test for English L2: A spoken receptive meaning-recognition test intended for Dutch-speaking L2 learners of English. *Language Teaching Research*. https://doi.org/10.1177/13621688221147462

Department for Education. (2013). *Languages programmes of study: Key stage 2*. National Curriculum in England. Retrieved from https://assets.publishing.service.gov.uk/media/5a7b9246e5274a7318b8f889/PRIMARY_national_curriculum_-_Languages.pdf.

Department for Education (2024). Schools, pupils and their characteristics. Retrieved from: https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics.

Dudley, A., Marsden, E., & Bovolenta, G. (2024). A context-aligned two thousand test: Towards estimating high-frequency French vocabulary knowledge for beginner-to-low intermediate proficiency adolescent learners in England. *Language Testing*. https://doi.org/10.1177/02655322241261415

Dunn, K. J. (2024). Random-item Rasch models and explanatory extensions: A worked example using L2 vocabulary test item responses. *Research Methods in Applied Linguistics, 3*(3). https://doi.org/10.1016/j.rmal.2024.100143

Edmonds, A., Clenton, J., & Elmetaher, H. (2022). Exploring the construct validity of tests used to assess L2 productive vocabulary knowledge. *System, 108*. https://doi.org/10.1016/j.system.2022.102855

Goldstein, H. (1982). Models for equating test scores and for studying the comparability of public examinations. *Educational Analysis, 4*(3), 107–118.

Graham, S., Courtney, L., Marinis, T., & Tonkyn, A. (2017). Early language learning: The impact of teaching and teacher factors. *Language Learning, 67*(4), 922–958. https://doi.org/10.1111/lang.12251

Hair, J., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Educational International.

Hasselgren, A. (2012). Assessing young learners. In G. Fulcher, & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 93–105). Routledge. https://doi.org/10.4324/9780203181287.

Jordan, E. (2012). Cognates in vocabulary size testing—A distorting influence? *Language Testing Asia, 2*(3), 5–17. https://doi.org/10.1186/2229-0443-2-3-5

Kasprowicz, R.E., Graham, S., & Morea, N. (forthcoming). Child-centred quantitative research. In Y.G. Butler, & A. Pinter (Eds.), *Child-focused approaches to applied linguistic research*. John Benjamins.

Kasprowicz, R.E., & Graham, S.(in progress). *Perceptions of primary languages education in England*. Manuscript in preparation.

Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching & Research, 4*(5), 900–913. https://doi.org/10.4304/jltr.4.5.900-913

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing, 21*(2), 202–226. https://doi.org/10.1191/0265532204lt277oa

Laufer, B., & Paribakht, T. S. (2008). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning, 48*(3), 365–391. https://doi.org/10.1111/0023-8333.00046

Leeming, P., & Harris, J. (2024). The language learning orientations scale and language learners' motivation in Japan: A partial replication study. *Research Methods in Applied Linguistics, 3*(1). https://doi.org/10.1016/j.rmal.2024.100096

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2010). When to stop removing items and persons in Rasch misfit analysis? *Rasch Measurement Transactions, 23*(4), 1241.

Linacre, J.M. (2023). Winsteps® Rasch measurement computer program User's Guide. Version 5.6.1. Winsteps Help for Rasch Analysis 5.7.2.

Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French: Core vocabulary for learners*. Routledge.

Mackey, A., Marsden, E., & Plonsky, L. (2016). The IRIS repository. In E. Marsden, & A. Mackey (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). Taylor and Francis. https://doi.org/10.4324/9780203489666.

McKay, P. (2006). *Assessing young language learners*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733093

McLachlan, A. (2009). Modern languages in the primary curriculum: Are we creating conditions for success? *Language Learning Journal, 37*(2), 183–203. https://doi.org/10.1080/09571730902928078

Milton, J. (2008). French vocabulary breadth among learners in the British school and university system: Comparing knowledge over time. *Journal of French Language Studies, 18*(3), 333–348. https://doi.org/10.1017/S0959269508003487

Morea, N., Kasprowicz, R. E., Morrison, A., & Silvestri, C. (2024a). *A new receptive vocabulary size test for young language learners in England. University of Reading*. University of Reading. Dataset. https://doi.org/10.17864/1947.001365

Morea, N., Kasprowicz, R. E., Morrison, A., & Silvestri, C. (2024b). *Receptive vocabulary size test for young language learners learning French, German or Spanish. Vocabulary test from "Diverse population, homogenous ability: The development of a new receptive vocabulary size test for young language learners in England using Rasch analysis. [Text/Language test]*. UK: IRIS Database, University of York. https://doi.org/10.48316/wGeqE-NS9al

Nation, I. S. P., & Anthony, L. (2016). Measuring vocabulary size. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 355–368). Routledge. III.

Nikolov, M., & Timpe-Laughlin, V. (2021). State-of-the-art article: Assessing young learners' foreign language abilities. *Language Teaching, 54*, 1–37. https://doi.org/10.1017/S0261444820000294

Phipps, J. (2023). The validation of two L2 self-efficacy instruments using Rasch analysis. *Research Methods in Applied Linguistics, 2*(3). https://doi.org/10.1016/j.rmal.2023.100084

Pignot-Shahov, V. (2012). Measuring L2 receptive and productive vocabulary knowledge. *Language Studies Working Papers, 4*, 37–45.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207–230.

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence, 67*, 44–66. https://doi.org/10.1016/j.intell.2018.01.003

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition, 15*(1), 157–166. https://doi.org/10.1017/S1366728910000623

Stæhr, L. S. (2008). Vocabulary size and skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139–152. https://doi.org/10.1080/09571730802389975

Tong, Y., Hasim, Z., & Halim, H. A. (2022). The impact of L2 vocabulary knowledge on language fluency. *Pertanika Journal of Social Sciences & Humanities, 30*(4), 1723–1751. https://doi.org/10.47836/pjssh.30.4.14

Tschirner, E., & Möhring, J. (2019). *A frequency dictionary of German: Core vocabulary for learners* (2nd ed.). Routledge.

Wallace, S. (2014). *A dictionary of education* (1st ed). Oxford University Press. https://www.oxfordreference.com/display/10.1093/acref/9780199212064.001.0001/acref-9780199212064.

Weng, F., & Liu, X. (2024). Exploring second language students' language assessment literacy: Impact on test anxiety and motivation. *Frontiers in Psychology, 15*. https://doi.org/10.3389/fpsyg.2024.1289126

Wind, S., & Hua, C. (2022). *Rasch measurement theory analysis in R*. Chapman & Hall.

Xing, H., Liu, N., Li, K., Cui, G., & Biering-Sørensen, F. (2024). Translation and validation of the Chinese self-report version of spinal cord independence measure (SCIM-SR): Rasch psychometric analysis and online application. *Computational and Structural Biotechnology Journal, 24*, 258–263. https://doi.org/10.1016/j.csbj.2024.03.029

Yamashita, T. (2022). Analyzing Likert scale surveys with Rasch models. *Research Methods in Applied Linguistics, 1*(3). https://doi.org/10.1016/j.rmal.2022.100022

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