**S2: Component skills of reading: tests of individual differences**

The battery of individual-difference tests in MECO L2 included the following measures:

(1) *Spelling Recognition Test* (adapted from Andrews & Hersch, 2010): In this test items are presented in a list and participants need to decide for each whether it is a correctly spelled word in English or not (i.e., mark each item as 'correct' or 'incorrect'). Half of the items are correctly spelled and the other half include spelling errors (e.g. seperate, benafit). The original version includes 88 items but given time considerations here we used a version with 44 items (22 correctly spelled and 22 spelling errors). Scores are the number of accurate responses across the 44 items: that is, the sum of the number of correctly spelled words marked by the participant as 'correct' and the number of incorrectly spelled words marked as 'incorrect'.

(2) *Vocabulary Knowledge Test* (adapted from Nation & Beglar, 2007). The original version of this task is a vocabulary size test of the receptive knowledge of English words selected from a frequency-ranked list of 14,000 English lemmas. Ten items are chosen from each 1000 words in the ranked list and are said to represent the respective frequency band. We refer to each 10 items from a frequency band as a 'thousand'. The original 14,000-word version of the original test contains 140 multiple-choice items (i.e., 14 thousands). Due to time considerations, here we used a shortened version with a maximum of 100 questions representing the top 10,000 words in the ranked frequency list of English. The test consists of a series of questions where a target word is embedded in a short non-defining context and participants need to choose its correct definition from four options. The 10 words and questions from each thousand were presented together on each screen. At the end of each group, a stopping rule was applied such that the test would stop if a participant had 4/10 incorrect answers or more. Two types of scores were calculated: The total number of correct responses across all completed word groups; and the total number of correct responses in thousands 2-5 (see main text for rationale).

Note that given the modifications applied to the test (i.e., trimmed frequency range; addition of stopping rule), our version does not measure an individuals' overall vocabulary size. Instead, we included it as a proxy of vocabulary knowledge, based on the premise that individuals with better vocabulary knowledge will know the definitions of more (and less frequent) words. We emphasize that the addition of the stopping rule does not restrict the validity of the test as a measure of vocabulary knowledge. Thus, as long as each item is positively correlated with the total expected score in the test (i.e., that the probability of answering correctly in each item is greater among readers with better overall vocabulary knowledge), and as long as questions in higher frequency bands are on average of higher difficulty than those in lower frequency bands, our stopping rule should not lead to a systematic bias in test scores (and indeed, both of these assumptions are supported by the Item Response Theory (IRT) analysis in Beglar, 2010). Further evidence for the validity of this test comes from its stronger correlation with LexTALE scores (*r* = 0.63, see Table 3 in the main text), which is a validated measure of vocabulary knowledge in L2 readers.

(3) *Motivation*. To assess the motivation of our participants to complete the study, we used the Student Opinion Scale (SOS) questionnaire (Thelk et al., 2009). SOS includes 10 statements that participants are asked to rank from ‘1=Strongly Disagree’ to ‘5=Strongly Agree’ according to how they feel about each of them in relation to completing the current study. Scores are the average rankings on the ten questions (after scale reversals when needed such that higher scores reflect higher motivation in all questions).

(4) *Lexical Test for Advanced Learners of English* (LexTALE; Lemhöfer & Boersma, 2012). A validated task aimed at measuring vocabulary knowledge in samples of medium to highly proficient L2 readers of English. It is an untimed lexical decision task, consisting of 60 trials: 40 words and 20 pseudowords. Scores are determined by the formula: ((number of words correct/40\*100) + (number of nonwords correct/20\*100)) / 2 (i.e., percentage of correct responses, corrected for the unequal proportion of words and nonwords).

(5) Test of Word Reading Efficiency – second edition (*TOWRE-II*; Torgesen et al., 2012). This task includes two subtests: Sight Word Efficiency (word reading) and Phonemic Decoding Efficiency (pseudoword reading). In each subtest participants are required to read aloud as many items as possible from a list of words/pseudowords within a 45-second time limit. Scoring in each subtest is based on the number of correctly read words/pseudowords.

In addition to measures from this battery of tests, we make use of two measures collected for all participants as part of the L1 portion of the study (reported in detail in Kuperman et al., 2020). These include scores the non-verbal IQ test from the Culture Fair Test-3 (CFT20, Subset 3 Matrices, short version, Form A, timed at 3 minutes, Weiß, 2006), and responses on an abridged version of the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian, et al., 2007). The former aimed at providing a comparable measure of non-verbal intelligence across sites, and the latter at collecting basic demographic and linguistic information about participants. For completeness, the data repository of MECO L2 includes the information collected from these measures as well.

**S3: Specifications of the eye-tracking laboratory setup.** Information regardingapparatus and reading task's settings.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **L1 of Sample** | **Eye-tracker** | **Font point size** | **Distance from screen** | **Characters per visual angle** | **Screen type** | **Screen resolution** | **Screen size** |
| Dutch | Eyelink 1000+ | 22 | 95 cm | 3.9 | Benq XL2420 | 1920 x 1080 | 24 inch |
| English | Eye Link 1000/1000+ | 20 | 60 cm | 3.3 | ThinkVision t24i-10 | 1920 x 1080 | 24 inch |
| Estonian | Eye Link 1000/1000+ | 20 | 85 cm | 2.08 | HP E233 | 1920 x 1080 | 23 inch |
| Finnish | Eye Link Portable Duo 1000 | 20 | 55 cm | 3.96 | Asus | 1920 x 1080 | 17 inch |
| German | Eye Link 1000+ | 20 | 60 cm | 3.3 | Alienware2518h | 1920 x 1080 | 25 inch |
| Greek | Eye Link 1000+ | 20 | 60 cm | 3.3 | BenQ XL2411P | 1920 x 1080 | 24 inch |
| Hebrew | Eye Link 1000+ | 20 | 80 cm | 1.97 | Asus VG248 | 1920 x 1080 | 24 inch |
| Italian | Eye Link 1000 | 20 | 95 cm | 2.65 | Acer | 1920 x 1080 | 27 inch |
| Norwegian | Eye Link 1000+ | 20 | 89 cm | 3.7 | BenQ XL2430-B | 1920 x 1080 | 24 inch |
| Russian | Eye Link 1000+ | 20 | 75 cm | 2.96 | ViewSonic Graphics Series G90fB (CRT) | 1600 x 1024 | 19 inch |
| Spanish | Eye Link 1000 | 20 | 70cm | 3.1 | LG 23MP55 | 1920 x 1080 | 23 inch |
| Turkish | Eye Link 1000+ | 20 | 60 cm | 3.3 | HP Pavilion 23cw | 1920 x 1080 | 23 inch |