Studying texts in a second language:
The importance of test type*

Heleen Vander Beken, Marc Brysbaert

Ghent University, Belgium

*This study was supported by a GOA grant from the Research Council of Ghent University (LEMMA Project). We would like to thank Dr. Michael Stevens for some exploratory data analyses and Dutch and English teacher Sarah De Paepe for rating files. In addition, we are grateful to Dr. Evelyne Lagrou for her contribution in the data collection. Finally, we wish to express our gratitude to three anonymous reviewers for their suggestions.

Address for correspondence: Heleen Vander Beken

Department of Experimental Psychology

Ghent University

Henri Dunantlaan 2

B-9000 Gent

Belgium

Tel. +32 9 264 94 27

Fax. +32 9 264 64 96

heleen.vanderbeken@ugent.be

Abstract

Little is known about the extent to which information encoding and retrieval differ between materials studied in first and second language (L1 and L2). In this study we compared memory for short, expository texts in L1 and L2, tested with a free recall test and a true/false judgement test. Our results show that students performed at the same level on the recognition test in both languages but not on the free recall test, with much lower performance in L2 than in L1, defined here as the dominant language. The L2 recall cost suggests that students' performance may be underestimated if they are exclusively tested with essay-type exams in L2.

The high mobility of students and the increased use of English as *lingua franca* in education mean that many people are taking courses in a language different from their native language. Surprisingly, little is known about how information studied in a second language (L2) is encoded in memory and to what extent retrieval differs from information learned in a first language (L1).

Bilingual language processing research has mainly focused on word recognition and word production. The general conclusion from this research is that both languages of a bilingual are active during language perception and production, even when only one language is needed (e.g. Van Assche, Duyck, & Hartsuiker, 2012). Less is known about how the meaning of words and texts is encoded in and retrieved from memory.

The general assumption among bilingualism researchers has been that meanings are stored as amodal, language-independent concepts and propositions, shared among the languages of a multilingual (for a review of the word recognition models, see Brysbaert & Duyck, 2010). Related to the issue of discourse and text representation in the brain, the same assumption goes back to studies in the 1960-1980s (e.g., Alba & Hasher, 1983; Sachs, 1967; Schank, 1972). The line of research started from the observation that people usually do not remember the specific wordings of a text (the surface structure) but recall the main ideas conveyed by it (the deep structure). This suggested that thought had a language of its own, in which content words were replaced by concepts, and the relationships between the words by a limited number of dependencies and causal chains between concepts. As Schank (1980, p. 244) summarized:

> … because people could easily translate from one language to another and, in a sense, think in neither, there must be available to the mind an interlingual, i.e., language-free, representation of meaning.

Within the view of language-independent thought representations, differences between L1 and L2 memory performance are explained by differences in translating the language input to thought representations and the thought representations to language output. On the word level, this assumption recurs in the asymmetry of connections between words and their meanings. Indeed, in the Revised Hierarchical Model (Kroll & Stewart, 1994), it is assumed that L2 words have weaker connections with their semantic concepts than L1 words, so that they sometimes have to activate their concepts via L2 → L1 translations. This is assumed to be particularly true for low levels of L2 proficiency.

Nott and Lambert (1968) published data in line with the model. They observed that bilingual participants recalled equal numbers of words in L1 and L2 when the words were presented in random lists, but not when the words could be organized into semantic categories. In the latter case, performance in L1 was better, unless the participants were told explicitly about the organisation of the list (in which case L2 performance again equalled L1 performance). The observation that participants were able to benefit more from semantic associations in L1 than in L2, agrees with the idea that access to semantic concepts is easier for L1 words. Still, it should be taken into account that the Revised Hierarchical Model and the research of Nott and Lambert (1968) involve the storage of individual words, which may differ from the storage of meaningful text materials.

Against the view of language-independent thought representations, there is some evidence that thought representations may not be completely language-independent (also see Alba & Hasher, 1983, for a review of the evidence that discourse memories may include more surface details than assumed by theories based on language-independent representations). First, autobiographical memory seems to be partially language-dependent. Memories of events are explained in more detail in the language in which the event took place and tend to differ depending on the language of the memory cues provided (Marian & Fausey, 2006;

Matsumoto & Stanny, 2006; Schrauf et al., 1998). Second, Watkins and Peynircioglu (1983) presented their participants with mixed lists of eight Spanish and eight English words. At the end of each list, participants were given word fragments, which they had to complete. Some of these word fragments were from a target word presented (e.g., *-lo--*for *cloud*), others were from the translation of a word presented (e.g., *-lo--* for the Spanish equivalent *nube*). If participants were unable to provide the correct word, more letters were added to the fragment (*-lo-d*; *clo-d*) until the participant was able to give the correct response. Watkins and Peynircioglu (1983) observed that more letter cues had to be given when the word fragments referred to translations than to the target words themselves, suggesting that the information stored in memory included more than language-independent semantic representations. A third piece of evidence was reported by Marian and Fausey (2006), who ran an experiment in which bilinguals were taught domain-specific information from auditory input in L1 or L2. Retrieval was more accurate and faster when the language of retrieval was the same as that of encoding, at least for highly-proficient bilinguals. Finally, multilinguals prefer to do arithmetic in the language used at school. Apparently, counting and tables of multiplication and addition are encoded in a language-specific way (Van Rinsveld, Brunner, Landerl, Schiltz, & Ugen, 2015).

The existence of language-dependent memory cues suggests that if such cues are present in the memory representations of texts, it may be more efficient to retrieve the information in the same language as the one used for learning.

To conclude: Psycholinguists thus far have done little systematic research on encoding and retrieving text information presented in L1 or L2 (see the discussion section for two small-scale studies). In their models of word processing, they assume the existence of language-independent meaning representations, to which the language input must be translated and which are translated again for verbal output, in line with ideas developed in the

1960-1980s. At the same time, there is some evidence that memories for text and discourse may be more accessible in the language studied than in another language mastered.

The reason why research on text memory in L1 vs. L2 has been so limited might be the complexity of the matter. Learning and remembering texts involve many variables, related to the learning materials, the learner, and the tests to be completed, so that any study answers only a fraction of the questions researchers and readers are likely to have.

For a start, many different types of texts can be chosen, even if the study is limited to printed materials. Texts can differ hugely in terms of contents (e.g., fiction vs. non-fiction), length (going from a 100-word paragraph to a 10,000 word chapter), and difficulty (both in terms of vocabulary, syntactic structures, background knowledge needed, and the number of inferences that must be made).

The most important learner-related variable is the L2-proficiency relative to the text difficulty. Information that is not understood can hardly be remembered. So, language proficiency in L2 (Droop & Verhoeven, 2003) and vocabulary knowledge (Cromley, Snyder-Hogan, & Luciw-Dubas, 2010; Mehrpour & Rahimi, 2010) are relevant. In addition, factors influencing reading comprehension must be considered. These include reading fluency (Başaran, 2013), prior knowledge (Coiro, 2011; Cromley et al., 2010), reading motivation (Andreassen & Bråten, 2009), working memory capacity (Conners, 2008; McVay & Kane, 2012), IQ (Keenan & Meenan, 2014), and strategy use (Cromley et al., 2010), among other variables.

Finally, the way in which memory is tested is likely to make a difference as well. Traditionally, a distinction is made between recognition and recall (e.g., Gillund & Shiffrin, 1984). Though both test types tap into declarative memory (Haist, Shimamura, & Squire, 1992), the processes of retrieval and the conditions for success differ (Hogan & Kintsch,

1971). Recall involves an extended search, which is slow and uncertain and which requires more processing resources, as is proven by a decreased recall performance (compared to recognition) with increasing age (Craik & Mcdowd, 1987). A recognition test includes many more cues, so that memory traces can be accessed more directly. In particular, true/false judgements can be considered as "locating questions" according to Guthrie (1998, see also Tal et al., 1994). Guthrie pointed out that the processes needed to locate details in a text are distinct from the processes involved in recalling the main ideas of the same text. Similarly, it may make a difference if one has to match a detailed ("locating") question to stored information than when one has to reproduce the core ideas from that same memory without cue. Alba and Hasher (1983) provided evidence that recall tests are more influenced by the participant's memory schemas and scripts than recognition.

Another way to conceive the difference between recognition and recall tests is to think of recognition tests as making it possible to probe for 'marginal knowledge', knowledge in memory that cannot be retrieved without the help of memory cues. Interestingly, probing for marginal knowledge via a recognition test may strengthen the memory trace to such an extent that it becomes available for recall. For instance, Cantor, Eslick, Marsh, Bjork, & Bjork (2014) reported that the administration of a multiple-choice test improved performance on subsequent recall test.

When confronted with such a multitude of potentially important variables, it is tempting to run a series of small experiments, addressing the various questions and possible confounds. A danger in doing so, however, is that each experiment tends to be underpowered, because of resource constraints. As has been well documented, this involves two risks. The first is that a null effect is obtained, which cannot be interpreted. The second is that a significant effect is found, which cannot be replicated (Gelman & Carlin, 2014), in particular when effects are close to the significance level (Francis, 2012; Leggett, Thomas, Loetscher, &

Nicholls, 2013; Simmons, Nelson, & Simonsohn, 2011). To avoid these problems, we ran a power analysis before setting up the experiment (see under Method).

Because of the importance of the test type, we decided to focus on this variable and compared a free recall test to a true/false judgement test in L1 and L2. Dutch-English bilinguals were asked to study a short text in their dominant language (L1) or in English (L2). Afterwards they either had to write down as much as they remembered from the text, or they had to answer a list of true/false questions about the text. To compare our findings to 'natural' studying, we used expository, factual texts. Since we assumed it is harder for participants to understand a text in L2, we expected lower results in English than in Dutch. We also expected a robust effect of test type, with the recognition test yielding higher results than the free recall test. We were particularly interested in the size of the L2 disadvantage to answer the practical question: are L2 education and examination so disadvantageous to students (d > .4; Ferguson, 2009) that they require remediation?

Method

Participants. To decide on the number of participants needed for a sufficiently powered experiment, we started from the observation that an effect size of d = .4 is seen as a practically significant effect. Such minimum effect size is usually required for efficient therapies and for group differences that must be addressed in applied settings (e.g., education; Ferguson, 2009). Since our design included a between-groups variable, we needed two groups of 100 participants to have 80% chance of detecting an effect of d = .4 (Cohen, 1992; see also Callens, Tops, & Brysbaert, 2012).

A total of 199 first year psychology students from Ghent University took part in the experiment in partial fulfilment of course requirements and for an additional financial reward (data collection was planned for 200 participants, but one student did not show up on any of the sessions they were invited to). All participants were Dutch native speakers who had studied English in high school for at least four years and who were regularly exposed to (subtitled) English television programs and English songs. In some of their university courses English handbooks were used, even though the teaching happened in Dutch. Note that, in this study, L1 was defined in terms of dominant language, not as the first acquired language. The data of four students who did not have Dutch as their dominant language were excluded from all analyses, so that the final analyses are based on N = 195. The participants' mean age was 18.6 yrs (sd 2.3); 129 were female students, 66 male. Participants were randomly assigned to the conditions.

Materials

Texts. We used two short, English texts from a study of Roediger and Karpicke (2006). Each text covered a topic in the domain of natural sciences: the Sun and sea otters. The English texts were slightly adapted for consistency. First, all spelling was altered to the US standard, to allow the use of consistent lexical measures such as word frequency. Second, culture-specific measurement units like 'inches' and 'pounds' were converted into the metric system the participants were familiar with, such as 'centimeters' and 'kilograms', terms that were used in the Dutch translation too. If these terms had not been changed in the English version, the difference between both language versions could have yielded a higher processing load in the English condition because Belgian students are not familiar with the American units.

The English texts were translated into Dutch. To check for ambiguous translations, they were then independently retranslated into English. If any semantic or syntactic ambiguity was found, we chose different translation equivalents to make the texts as similar as possible in both languages. All content words were matched between languages for total word form frequency and word form frequency for the specific part of speech. Frequencies were taken from SUBTLEX-US (Brysbaert & New, 2009) and SUBTLEX-NL (Keuleers, Brysbaert, & New, 2010). They were transformed to Zipf-values as a standardised measure to account for different corpus size (van Heuven, Mandera, Keuleers, & Brysbaert, 2014). No absolute criterion was used, but when frequencies differed by more than one Zipf unit, a Dutch synonym was selected that matched the English frequency more closely. In Dutch, the number of compound nouns is inherently higher, so the same concept is often presented by a compound noun in Dutch and by a combination of nouns in English. For matching purposes in these cases, the compound word frequency in Dutch was compared to the word bigram frequency in English. The same technique was used when certain fixed expressions or phrasal verbs differed inevitably between languages. This resulted in one English text about the Sun, 258 words long, with a Dutch translation of 248 words, and one English text about sea otters, 279 words long, with a Dutch translation of 274 words. Welch two sample t-tests comparing the word frequency distributions between the English and Dutch version of *The Sun* indicated that both texts were comparable ($t(488) = 0.94$, $p > .250$). The same was true for the two texts on sea otters ($t(527) = -0.19$, $p > .250$).

The texts were presented on paper in Times New Roman 12, as in Roediger and Karpicke (2006). Line spacing was 1.5 and the first line of every paragraph was indented.

Free recall and true/false judgement tests. Two types of tests were administered to accompany the texts: a free recall test and a true/false judgement test. In the free recall test, participants received the following instruction: “Write a summary of the text you have just

read. Be as detailed as you can be". This way, participants were not asked to literally reproduce the text, but to produce the ideas and we encouraged them to add details when possible.

Roediger and Karpicke (2006) divided their texts into 30 ideas or propositions that had to be reproduced. This list (with adaptations analogous to the text adaptations) was used as a scoring form for the free recall tests in English, and a Dutch version was created. Next, a true/false test of 46 questions was developed for both texts. 30 true/false questions were derived from the ideas on the free recall scoring form. Those questions were literal questions in which one concept was slightly changed for items that require a FALSE response. For example: "The Sun today is a white dwarf star" requires a FALSE response since the text states that "The Sun today is a yellow dwarf star". Next, 10 inferential questions were written (see Tal, Siegel, & Maraun, 1994 for a study on similar question types), half of which were based on one proposition in the text, and half of which were based on several propositions from several locations in the text, requiring the integration of ideas. An example of such a question is "The surface of a red giant star is hotter than that of a yellow dwarf star". To respond to that question, the reader has to remember and integrate information about the surface temperature of two of the mentioned star types. In addition, 6 false memory questions were created containing a statement that was not mentioned in the text but was in some way related to a concept in the text. An example of such a statement is "Sea otters live around Alaska", while Alaska was mentioned in the text as the location of an oil spill but not described as sea otters' necessary habitat. All questions were translated to Dutch. For this test, the instruction was "Tick the correct answer box for every statement, based on the text you have just read". Instructions for the tests were written on the test form itself, in the language of the test. All tests were administered on paper.

To make sure that the questions from the true/false test could not be answered on the basis of prior knowledge, we administered the statements to a pilot group of 38 participants similar to the group tested in the experiment, and asked them to complete the true/false test to the best of their knowledge (see Coleman, Lindstrom, Nelson, Lindstrom, & Gregg, 2010 for an example of such a *passageless administration* post-test on a widely used reading comprehension test). This passageless administration indicated that the scores were slightly higher than the expected 50% for both *The Sun* (M = 55%, [range of correct answers to questions across participants: 41%-70%]) and *Sea Otters* (M = 55% [41%-65%]). Therefore, the questions were analysed individually. If the results were significantly above chance level for a certain question, the question was excluded from the test. A one-tailed binomial test with a Dunn-Šidák correction for multiple testing (46 statements) indicated that, for *The Sun,* 5 questions were answered significantly better than chance, and 9 questions for *Sea Otters*. When these questions were excluded, the means decreased to M = 52% [34% - 68%] for *The Sun* and to M = 47% [35%- 57%] for *Sea Otters.* These questions were excluded from the analysis of the main experiment, resulting in 41 true/false items for *The Sun* and 37 for *Sea Otters.* The texts and the tests can be obtained from the authors for research purposes.

Motivation and Text-related Questionnaires. After the true/false and free recall tests, the participants completed two questionnaires. The first asked about their general attitude towards reading and testing: their testing motivation, their self-perceived level of performance relative to fellow students, and their general reading motivation in Dutch (L1) and English (L2). The second questionnaire checked for prior knowledge about the texts, the perceived difficulties (both content and structure) of the texts, and how interesting the texts were. The questionnaires were presented in Dutch to all participants, using 7-point Likert scales.

Subjective assessment of language proficiency. The participants' language background was assessed with a Dutch version of the Language Experience and Proficiency Questionnaire

(Marian, Blumenfeld, & Kaushanskaya, 2007; translated by Lisa Vandeberg; adaptation Freya De Keyser, Ghent University, and Marilyn Hall, Northwestern University).

Objective L1 proficiency tests. L1 proficiency was measured with the Dutch LexTALE test, a language-specific lexical decision test containing 40 words of various difficulty levels and 20 nonwords (Lemhöfer & Broersma, 2012). In addition, the participants received a semantic vocabulary test in a multiple choice format with four answer alternatives and a Dutch spelling test in which they had to spell words ofvarious spelling difficulties that were read aloud (all developed at the department).

Objective L2 proficiency tests. L2 proficiency was measured with the English LexTALE test of vocabulary knowledge for advanced learners of English (Lemhöfer & Broersma, 2012). Next, the participants received a version of the MINT picture-naming task, adapted for Dutch-English speakers (Gollan, Weissberger, Runnqvist, Montoya, & Cera, 2012), in which they saw a black- and white picture of an object of which they had to type the English name. The Oxford Quick Placement Test (QPT; 2001) was also administered, which is considered a measure of general proficiency consisting of multiple choice items of vocabulary and sentence comprehension and grammar (verb use, part of speech regulations, …). Finally, an English spelling test was given, similar to the Dutch spelling test (developed at the department).

Measures of reading exposure, intelligence and WM. A Dutch author recognition test (modelled after Moore & Gordon, 2015) was used to estimate the participants' familiarity with authors' names, and thus the time they spend reading and acquiring language skills. Intelligence was measured with the Raven Progressive Matrices (short version, Bors & Stokes, 1998), and working memory with the automated operation span task, which provides a measure of working memory capacity (Unsworth, Heitz, Schrock, & Engle, 2005).

Distractor task between learning and testing. A computerized version of the Corsi block-tapping task (Corsi, 1972) with English instructions was used as a distractor task between every study phase and test phase. A similar distractor task was used by Roediger and Karpicke (2006). They asked their participants to solve multiplication problems for two minutes. We opted for a Corsi-task because research has shown that arithmetic fact retrieval, especially multiplication, is related to phonological processing (De Smedt & Boets, 2010), which would have activated the L1 of our participants. We wanted to avoid this strong internal L1-activation. The Corsi task is a visuo-spatial short-term memory test and requires the participants to repeat sequences indicated on an array of blocks. The test begins with a short sequence and increases until the participant makes too many mistakes. Since the general instructions of the experiment were in Dutch and the Corsi-task instructions were in English, both L1 and L2 were shortly activated for both language-groups, cancelling out pre-activation effects of one language.

Procedure. Participants were assigned to one of eight conditions: two language groups that were further divided into four conditions in which the text order and the test type order were counterbalanced, to make sure that the results were not confounded by any of the control variables (2 x 2 x 2 factorial design). This is illustrated in Figure 1.

	Dutch				**English**			
Study phase I	De zon		Zeeotters		The Sun		Sea Otters	
Test phase I	True/false judgment	Free recall	True/false judgment	Free recall	True/false judgment	Free recall	True/false judgment	Free recall
Study phase II	Zeeotters		De zon		Sea Otters		The Sun	
Test phase II	Free recall	True/false judgment	Free recall	True/false judgment	Free recall	True/false judgment	Free recall	True/false judgment

Fig. 1 The eight experimental conditions to which all participants were randomly assigned.

Tests were administered in groups of 50 participants at most. Oral instructions were given in Dutch. Participants were told to follow the instructions for each part of the experiment and to wait for new instructions before advancing to the next task. They were informed that they had to study a text within a limited time frame and that they would be tested for their knowledge, but not with what type of test. Texts and tests were presented on paper. Participants studied the first text passage for seven minutes. Next, they took part in the computerized Corsi-task. The participants were asked to interrupt the task after two minutes and start a 7-minute test period in which they had to take the first recall or true/false judgement test. They were not allowed to look back at the text. After the 7-minute testing phase, the full procedure was repeated for the second text. The language of the texts and tests remained constant but the test type was changed (i.e., participants did both the recall and the true/false test in L1 or in L2).

After the second test, participants filled in the various questionnaires and completed the language and IQ tests. The English and the Dutch LexTALE, the Dutch semantic vocabulary test, and the Oxford QPT were administered individually online; all other tests were administered during the group sessions. The experiment took two hours in total.

Results

S*coring*. In our marking of the free recall tests we followed the guidelines set out by Roediger and Karpicke (2006). We scored the presence and correctness of the ideas from the text, irrespective of spelling errors and the overall organization of the recall protocol. Participants received 1 point for every correctly recalled idea and 0 points if the idea was recalled incorrectly or not recalled at all. If an idea was partially recalled, a .5 score was given. For the text about sea otters, three propositions had to be split into two separate ideas,

because often only one of them was recalled. A random sample of 100 recall forms was scored by two raters: the first author and a Dutch-English teacher with test rating experience. The second rater got the following guidelines: Spelling and grammatical mistakes must not be punished unless those mistakes obscure meaning (a similar guideline is given for the PISA tests; see appendix in Cartwright, 2012). We first calculated the interrater reliability: the Pearson correlation between the scores of both raters across all forms was .85. Partial analyses showed similar results of r = .83 for the Dutch ratings only, r = .87 for the English texts, r = .83 for *The Sun* and r = .86 for *Sea Otters*. Given the reassuring correlation, the rest of the tests were rated by the experimenter only. Since both raters barely ever used the .5-score (33/3210 trials) and did not agree with each other on those scores, we replaced them by 0.

The true/false judgements were scored dichotomously (correct/incorrect) with a correction key. After exclusion of the questions that came out poorly in the passageless administration test, we calculated the percentages of correctly answered questions and calculated percentages of correctly recalled ideas for the free recall test as well.

All data are available at https://osf.io/2twzd/ (Open Science Framework).

Testing whether the students were matched in the L1 and L2 condition. Because the main comparison involves L1 vs L2 studying, we first checked whether both groups were matched on the control variables we assessed. Table 1 and 2 show that this was the case. There were no significant differences between the two groups if a correction for multiple testing was taken into account[1]. In addition to this group comparison, we looked into within-subject differences in motivation of the total group of participants. Interestingly, participants in general had a higher reading motivation in L1 (M = 5.18, SD = 1.41) than in L2 (M = 4.51, SD = 1.47; Wilcoxon signed rank test resulted in V = 8610.5, $p < .001$). The reliability of the measures was measured using Cronbach's alpha, which was generally high. Only the Raven's

matrices resulted in an alpha of .47, probably due to an error in the administration (we presented each question on a central screen for the same duration, in group, while normally, the test is taken individually and participants can move through the items at their own pace). Table 3 displays the reliability measures and the correlations between the various measures.

< Insert Table 1 about here >

< Insert Table 2 about here >

< Insert Table 3 about here >

Assessing the participants' L2 proficiency level. The performance on various tests allowed us to assess the L2 proficiency level of the participants. Table 1 shows that the scores on the English LexTALE (M = 72) were much lower than those on the Dutch LexTALE (M = 89). Lemhöfer and Broersma (2012) reported scores of M = 75 on the English LexTALE for Dutch-English students in the Netherlands and M = 65 for Korean-English bilinguals. Elgort, Candry, Eyckmans, Boutorwick, and Brysbaert (in press) observed scores of M = 75 for a group of students similar to the one tested here, and M = 44 for a group of Chinese-English bilinguals, who were either pre-degree or in the first year of an undergraduate degree at a New Zealand university. Cop, Dirix, Drieghe, and Duyck (2016) reported scores of M = 91 for English native speakers and M = 76 for a group of Dutch-English bilinguals very similar to the participants we tested.

A score of 44 on the QPT places the participants in the upper intermediate band of that test. Lemhöfer and Broersma (2012) reported scores of M = 46 for their Dutch-English bilinguals and M = 38 for the Korean-English bilinguals.

All in all, the bilinguals we tested were unbalanced bilinguals with a reasonably good command of English, in line with what could be expected on the basis of their high school studies and the language demands placed on them at university.

Performance on the memory tests. To analyse memory performance, we used a 2 (language group) x 2 (test type) mixed ANOVA. Participants had been divided in eight groups, each presented with only one language and one combination of text and test type in a counterbalanced order. Given that the texts and presentation orders were control variables, to be counterbalanced, they were not included in the analysis[2]. As each participant did one recognition and one recall test, this was a repeated measure. The analysis indicated a significant main effect of language ($F(1,193) = 19.88$, $p < .001$, $\eta^2_p = .09$), a significant main effect of test type ($F(1,193) = 286.79$, $p < .001$, $\eta^2_p = .59$, Type III Anova) and, most importantly, a significant interaction between both variables ($F(1,193) = 30.25$, $p < .001$, $\eta^2_p = .14$). Figure 2 shows the effects (see also Table 4 for the exact data). Separate comparisons indicated that the difference between L1 and L2 was not significant for the recognition test (Cohen's $d = .07$; $F(1,193) = .26$, $\eta^2_p = .001$)[3], but resulted in a large effect size for the recall test ($d = .86$; $F(1,193) = 35.68$, $p < .001$, $\eta^2_p = .16$). Cohen's d values of .2 are usually considered "small", .5 "medium" and .8 or higher as "large". As indicated in the method section, d values of .4 and more are considered to be of practical relevance in applied settings.

Since the true/false test had only two response alternatives, it had an estimated minimal performance level of 50% (chance level), hampering the comparison of the recognition test with the recall test.[4] A simple equation to correct for this, is to recode the obtained recognition scores with the equation corrected_score = (raw_score – chance_score) / (maximum_score – chance_score). Applied to the yes/no test, a raw score of .80 results in a

corrected score of (.80 - .50) / (1.00 - .50) = .60 (or 60%). In such an analysis it is customary to level all scores under 50% to zero performance. When the analysis was redone with the corrected scores, the same pattern of results was obtained. The main effect of language remained significant ($F(1,193) = 11.14$, $p < .01$, $\eta^2_p = .05$), as did the main effect of test type ($F(1,193) = 7.12$, $p < .01$, $\eta^2_p = .04$, Type III Anova). Most importantly, the interaction between both variables remained significant ($F(1,193) = 12.76$, $p < .001$, $\eta^2_p = .06$). A separate comparison indicated that the difference between L1 and L2 was not significant for the recognition test (Cohen's d = .08; $F(1,193) = .29$, $\eta^2_p = .001$). Of course, the effect remained the same for the recall test, as this variable was not altered (d = .86; $F(1,193) = 35.68$, $p < .001$, $\eta^2p = .16$).

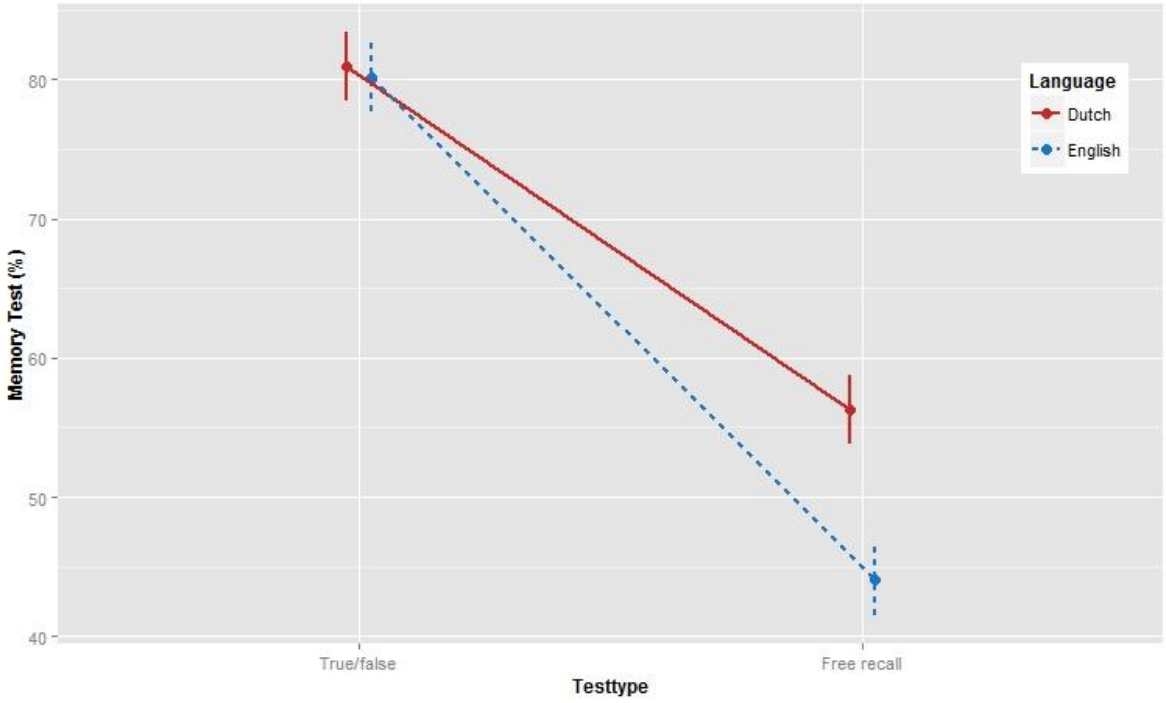

Fig. 2 Mean percentage of recalled ideas in all conditions with 95% confidence intervals. Note that chance level for a true/false test equals 50%. So the average scores on this test could be compared to a 60% score for the free recall test.

< Insert Table 4 about here >

Discussion

In this experiment, we tested how much information students remembered from short, expository texts studied in L1 and in L2. Two test types were compared: free recall and true/false recognition.

The free recall test measured how much students remember without being helped by memory cues. The recall processes assessed with such a test are very similar to those evaluated with open exam questions or essay-type exams. Because the goal of our study was to know how much information the participants could recall from the study materials independent of their L2 writing skills, we adopted the guideline (from PISA and other contexts) not to take into account spelling errors and grammatical mistakes in scoring the tests.

The recognition test was a test to measure as much knowledge as possible, including marginal knowledge, as defined by Cantor et al. (2014). Participants were given detailed statements and asked whether those were true or false according to the text.

Against our expectations, students did not show any difference in performance on the true/false test as a function of the language in which they had studied the text. They were correct on 80% of the questions (corresponding to a 60% score if corrected for guessing), both when they had studied in L1 or in L2 (Figure 2). This suggests that students understood the study materials equally well in L2 and L1 and did not perform at ceiling level.[5]

In contrast, participants studying in L2 performed significantly worse on the free recall test (44%) than the participants studying in L1 (56%). The difference corresponds to a large

standardized effect size of d = .86, meaning that it is of practical relevance in applied settings. In the remainder of the text we will call this difference the L2 RECALL COST. Because of the large number of participants we tested and the many precautions we took to make sure that both groups were matched, we can have confidence in the reliability and the replicability of the effect.

The L2 recall cost, together with the equivalent performance in the recognition test (result observed with the same participants), suggests that the cost is not simply due to deficiencies in the initial reading stage such as word encoding difficulties. In that case, we should have found lower performance on the L2 recognition test as well.

If word encoding is unlikely to be the origin of the L2 recall cost, we have to look for other factors. One of these may be that students are less able to express their thoughts in L2 writing. Their understanding is the same for texts studied in L1 and L2, but they have an L2 recall cost because they experience difficulties in translating thoughts into written L2 output, either as a consequence of weaker L2 writing skills in general or of weaker L2 retrieval.

An interesting idea in this respect is that it may be possible to train the translation of thoughts into L2 output. Karpicke and Roediger (2007) observed that their students remembered more from a text after having taken a test than after been given the opportunity to study the text for a second time. One of the explanations they proposed for this "testing effect" was that taking a test provided students with practice in retrieval processes. If this explanation is valid, we may be able to diminish the L2 recall cost by providing L2 students with practice in L2 recall before they take a test (or exam). In this respect, it may also be of importance that our participants did not know beforehand which test they were getting. It may be that students study differently if they know they will have to take a written essay-type exam in L2.

Another reason for the L2 recall cost may be that L2 recall induces more stress. De Quervain, Roozendaal, Nitsch, McGaugh, and Hock (2000) found that raising the cortisol level by administering cortisone impaired recall but not recognition of a word list. The effects were not found when the cortisone was administered before encoding, indicating that the impairment was associated with the retrieval process instead of the encoding process.

A way to test whether the L2 recall cost is related to difficulties in expressing one's knowledge in L2 (rather than to the knowledge itself), is to have participants learn a text in L2 but test them in L1. According to Joh (2006), several authors have suggested that L2 testing is disadvantageous for students because of limitations in expressing themselves in L2 (Wolf, 1993; Joh, 1998; Lee, 1986; as reported by Joh, 2006). For that reason, Joh (2006) interviewed his students in L1 even though the study was about L2 text studying (see Brantmeier, 2005 for another example).

Two other studies are relevant in this respect. Chen and Donin (1997) asked 36 Chinese-English bilinguals to read a short biology text in either Chinese or English, using a cross-lingual design with L2-L2, L2-L1 and L1-L1 conditions. Half of the students were biology students with high background knowledge of the topic; half were engineer students with a limited background. Participants were asked at four places within each passage to orally recall what they had just read, and they were asked to give detailed recall of the whole passage at the end of each text. The quality of recall differed as a function of the background knowledge but, surprisingly, it did not differ significantly between the language conditions tested, contrary to what we found. However, the condition L2-L2 with L2 recall seems to show a trend towards lower performance than the L2-L1 condition. The study may have lacked the power to pick up the difference. In addition, participants spent more time to read the text in L2 (remember we had a fixed studying time of 7 min).

Longer reading times were also reported by Donin et al. (2004). They asked 16 Canadian army officers to read English (L1) or French (L2) texts and to retell in English what they had read after every 4 sentences and after the full text. The participants needed more time to read in L2 than in L1, but memory accuracy after reading was equivalent. Again, however, the power of the study was very low.

Based on the studies mentioned above, the recall cost might be reduced if participants are allowed to take the recall test in L1 rather than in L2. At the same time, we must keep in mind that learning in L2 and testing in L1 involves a language change, which may harm performance if the memory representation of a text is not completely language-independent. Indeed, Marian and Fausey (2006) reported that for their spoken stimulus materials and their participants' retrieval was more accurate and faster when the language of retrieval was the same as that of encoding.

So far, we have assumed that the L2 recall cost is entirely due to difficulties in translating thought into L2 output. There are reasons, however, to believe that this may not be the correct or entire explanation for the L2 recall cost. It could be that the memory representation of a text read in L2 is less rich and organized than that of a text read in L1. A possible explanation may be found in van den Broek, Young, Tzeng, and Linderholm's (1999) Landscape model. According to these authors, a text is translated into a mental model consisting of a network of interrelated concepts (in this case, propositions and domain-specific content words). Factors like background knowledge and attention play a role in how concepts and their relations are placed in the mental model. During reading, the activation of concepts and their relations is continuously updated, resulting in a dynamic "landscape" of activation. Importantly, Van Den Broek et al. (1999, p. 77) also state that the processing of a concept is accompanied by cohort activation:

> When a concept is activated, other concepts that are connected to it [...] will be somewhat activated as well.

If we assume that the cohort of co-activated concepts is larger in L1 than in L2, we may have a mechanism that explains why the mental model of a text read in L1 is richer than that of a text read in L2. This accords with the word list recall findings of Nott and Lambert (1968) we discussed in the introduction. These authors found that semantic categorisation of lists helps memory more in L1 than in L2, in particular when the organisation is not made explicit. In addition, despite having a decent general understanding of the text, students might be unfamiliar with some of the domain-specific vocabulary (e.g. "badger"), resulting in less cohort activation from and to this concept. In the terms of Cantor et al. (2014), they might have 'marginal knowledge' of those propositions which are harder for them to understand, which means they can recognise the propositions, but cannot recall them. We also note that the participants in our study reported they were less motivated to read a text in L2 than in L1, which may have influenced the richness of the mental model they built.

A poorer mental model would also explain why the participants did not experience an L2 cost in the recognition test, as recall depends much more on the organisation of the mental model than recognition (Alba & Hasher, 1983). In this respect we have come to notice that the theoretical separation between semantic and episodic memory may not adequately reflect reality. Memory researchers typically make a distinction between semantic memory and episodic memory, with semantic memory being defined as consisting of general knowledge about the world and concepts, and episodic memory defined as dealing with episodes occurring in a given place at a given time. It can be questioned to what extent studying a text for a test (or an exam) results in semantic knowledge or episodic knowledge. As Van Den Broeck et al (1999, p. 80) point out:

> the modifications in semantic memory caused by a single text are likely to be small, […] unless a concept or set of concepts receives massive and/or repeated attention.

If text studying mainly results in episodic representations ("according to the text studied then and there, I have to answer that …"), then the type of memory representations we are studying may not be so much different from those studied in autobiographical memory (Marian and Fausey, 2006; Matsumoto & Stanny, 2006; Schrauf et al., 1998).

All in all, there are many possible explanations for our finding that free, written recall of an L2 domain-specific expository text has a cost for students, while there is no such L2 cost in a true/false judgement test for the same materials and participants. These are all avenues for further research. For instance, in future research it may be worthwhile to better examine the participants' production skills in L2, rather than the perception skills that were central in our current testing. Other questions that must be answered relate to the issues we mentioned in the introduction about the external validity of our finding: To what extent does the recall cost generalize to other types of texts, other learners, and other memory tests? We made a strong effort to ensure internal validity (so that we can rely on the data we observed) at the expense of the number of studies we were able to run.

Although our findings raise a list of theoretical issues (some of which we hope to address in the future, and some of which we hope others will find interesting to tackle), they do point to an important practical implication. The observation that students have a serious L2 recall cost and at the same time good L2 recognition performance, raises the question of what type of test they should be given for their exams. If all exams are essay-type exams, it is to be feared that L2 students will be at a serious disadvantage to obtain good grades (unless training helps them to acquire these skills and such training is offered to the students before they have to take their exams). On the other hand, if all exams are of the recognition type, L2 students

may find themselves even less able to talk about their knowledge (in L2). Much here, of course, depends on the type of skills taught in the course. If the skills are verbal, it can be defended that students should be able to express themselves in the language of their study. However, the situation becomes more complicated for less verbal skills. An L2 engineering student, for instance, may have learned perfectly how to design a machine, but not be able to explain this at the same level in L2 essay writing.

Given that we are dealing with a large effect size, these are issues we think education authorities will have to address, now that an increasing number of students are taught and tested in a language other than their native language.

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93*(2), 203–231.

Andreassen, R., & Bråten, I. (2009). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading, 33*(3), 263–283.

Başaran, M. (2013). Reading Fluency as an Indicator of Reading Comprehension. *Educational Sciences: Theory & Practice, 13*(4), 1–14.

Bors, D. A., & Stokes, T. L. (1998). Raven's advanced progressive matrices: norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*(3), 382–398.

Brantmeier, C. (2005). Effects of Reader's Knowledge, Text Type, and Test Type on L1 and L2 Reading Comprehension in Spanish. *The Modern Language Journal, 89*, 37–53.

Brysbaert, M., & Duyck, W. (2010). Is it time to leave behind the Revised Hierarchical Model of bilingual language processing after fifteen years of service? *Bilingualism: Language and Cognition, 13*(3), 359–371.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–90.

Cantor, A. D., Eslick, A. N., Marsh, E. J., Bjork, R. a., & Bjork, E. L. (2014). Multiple-choice tests stabilize access to marginal knowledge. *Memory & Cognition, 43*, 193–205.

Cartwright, F. (2012). Technical feasibility of Reporting YITS 2010 Skill Assessment Results on the PISA 2000 Reading Scale. *OECD Education Working Papers, 69*.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112(1),* 155-159.

Coiro, J. (2011). Predicting Reading Comprehension on the Internet: Contributions of Offline Reading Skills, Online Reading Skills, and Prior Knowledge. *Journal of Literacy*

Research : A Publication of the Literacy Research Association, *43*(4), 352–392.

Coleman, C., Lindstrom, J., Nelson, J., Lindstrom, W., & Gregg, K. N. (2010). Passageless Comprehension on the Nelson-Denny Reading Test: Well Above Chance for University Students. *Journal of Learning Disabilities*, *43*(3), 244–249.

Conners, F. a. (2008). Attentional control and the Simple View of reading. *Reading and Writing*, *22*(5), 591–613.

Cop, U., Dirix, N., Drieghe, D. & Duyck, W. (2016). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research methods*, 1-14.. doi:10.3758/s13428-016-0734-0

Corsi, P. M. (1972). Human memory and the medial temporal region of the brain. *Dissertation Abstracts International*, *34*(2), 891B.

Craik, F. I. M., & Mcdowd, J. M. (1987). Age Differences in Recall and Recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 474–479.

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. a. (2010). Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *102*(3), 687–700.

de Quervain, D. J.-F., Roozendaal, B., Nitsch, R. M., McGaugh, J. L., & Hock, C. (2000). Acute cortisone administration impairs retrieval of long-term declarative memory in. *Nature Neuroscience*, *3*(4), 313–314.

De Smedt, B., & Boets, B. (2010). Phonological processing and arithmetic fact retrieval: Evidence from developmental dyslexia. *Neuropsychologia*, *48*(14), 3973–3981.

Donin, J., Graves, B., & Goyette, E. (2004). Second Language Text Comprehension: Processing within a Multilayered System. *The Canadian Modern Language Review*, *61*(1), 53–76.

Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and

second-language learners. *Reading Research Quarterly*, *38*(1), 78–103.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*(6), 585–594.

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *9*, 1–67.

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of Spoken Language Dominance: A Multi-Lingual Naming Test (MINT) and Preliminary Norms for Young and Aging Spanish-English Bilinguals. *Bilingualism (Cambridge, England)*, *15*(3), 594–615.

Haist, F., Shimamura, A. P., & Squire, L. R. (1992). On the Relationship Between Recall and Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 691–702.

Hogan, R. M., & Kintsch, W. (1971). Differential Effects of Study and Test Trials on Long-Term Recognition and Recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 562–567.

Karpicke, J. D., & Roediger, H. L. I. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162.

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities*, *47*(2), 125–35.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650.

Leggett, N. C., Thomas, N. a, Loetscher, T., & Nicholls, M. E. R. (2013). The life of p: "Just significant" results are on the rise. *Quarterly Journal of Experimental Psycholog*, *66*(12), 2303–2309.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*, 325–343.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(August), 940–967.

Marian, V., & Fausey, C. M. (2006). Language-dependent memory in bilingual learning. *Applied Cognitive Psychology*, *20*, 1025–1047.

Matsumoto, A., & Stanny, C. (2006). Language-dependent access to autobiographical memory in Japanese-English bilinguals and US monolinguals. *Memory (Hove, England)*, *14*(3), 378–390.

McVay, J. C., & Kane, M. J. (2012). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *Journal of Experimental Psychology. General*, *141*(2), 302–20.

Mehrpour, S., & Rahimi, M. (2010). The impact of general and specific vocabulary knowledge on reading and listening comprehension: A case of Iranian EFL learners. *System*, *38*(2), 292–300.

Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: item response theory analysis of the author recognition test. *Behavior Research Methods*, *47*(4), 1095–1109.

Nott, C. R., & Lambert, W. E. (1968). Free Recall of Bilinguals. *Journal of Verbal Learning and Verbal Behavior*, (7), 1065–1071.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests

improves long-term retention. *Psychological Science*, *17*(3), 249–55.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Attention, Perception, & Psychophysics*, *2*(9), 437–442.

Schank, R. C. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, *3*(4), 552–631.

Schrauf, R. W., & Rubin, D. C. (1998). Bilingual Autobiographical Memory in Older Adult Immigrants: A Test of Cognitive Explanations of the Reminiscence Bump and the Linguistic Encoding of Memories. *Journal of Memory and Language*, *39*, 437–457.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366.

Tal, N. F., Siegel, L. S., & Maraun, M. (1994). The role of question type and reading ability in reading comprehension. *Reading and Writing*, *6*(4), 387–402.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505.

Van Assche, E., Duyck, W., & Hartsuiker, R. J. (2012). Bilingual word recognition in a sentence context. *Frontiers in Psychology*, *3*(June), 174.

van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The Landscape Model of Reading: Inferences and the Online Construction of Memory Representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The Construction of Mental Representations During Reading* (p. 404). Lawrence Erlbaum Associates.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology (2006)*, *67*(6), 1176–90.

Van Rinsveld, A., Brunner, M., Landerl, K., Schiltz, C., & Ugen, S. (2015). The relation

between language and arithmetic in bilinguals: insights from different stages of language acquisition. *Frontiers in Psychology*, *6*.

Watkins, M. J., & Peynircioglu, Z. F. (1983). On the Nature of Word Recall : Evidence for Linguistic Specificity. *Journal of Verbal Learning and Verbal Behavior*, *22*, 385–394

Table 1. *Mean scores of the language groups on the various proficiency and intelligence tests (standard devations between brackets).*

Tests	**L2 group (n = 97)**	**L1 group (n = 98)**	**All (n = 195)**
Gender	58F/39M	71F/27M	129F/66M
Age	18.39 (1.42)	18.82 (3.04)	18.61 (2.34)
Dutch LexTALE (max = 100)	89.52 (5.79)	89.31 (5.68)	89.42 (5.72)
Dutch vocabulary MC (max = 60)	42.03 (4.75)	41.70 (4.45)	41.87 (4.59)
Dutch spelling (max = 100)	78.25 (9.52)	79.06 (8.27)	78.71 (8.90)
English LexTALE (max = 100)	72.85 (10.95)	71.08 (9.08)	71.96 (10.07)
English Spelling (max = 100)	50.52 (17.82)	51.31 (14.35)	50.92 (16.14)
MINT (max = 60)	25.58 (11.64)	26.95 (11.81)	26.27 (11.71)
QPT (max = 60)	44.47 (6.63)	43.59 (6.58)	44.03 (6.61)
Author recognition (%hits - %false alarms)	26.09 (15.31)	22.61 (13.10)	24.34 (14.31)
Raven (IQ) (max = 12)	4.47 (1.96)	5.02 (1.82)	4.75 (1.90)
Operation Span (WM) (max = 75)	57 (13.42)	60.07 (12.07)	58.55 (12.81)

Note: The test statistics can be found at https://osf.io/2twzd/.

Table 2. *Mean scores of the language groups on the self-ratings included in the questionnaire (standard deviations between brackets).*

Self-ratings	**L2 group**	**L1 group**	**All**
Dutch speaking (max = 10)	9.49 (0.63)	9.33 (0.77)	9.42 (0.70)
Dutch comprehension (max = 10)	9.54 (0.62)	9.54 (0.67)	9.54 (0.64)
Dutch reading (max = 10)	9.46 (0.71)	9.45 (0.78)	9.46 (0.74)
English speaking (max = 10)	7.30 (1.04)	6.87 (1.35)	7.08 (1.22)
English comprehension (max = 10)	8.22 (1.13)	7.98 (1.43)	8.10 (1.29)
English reading (max = 10)	7.71 (1.35)	7.57 (1.25)	7.64 (1.30)
Dutch reading motivation (max = 7)	5.16 (1.54)	5.20 (1.30)	5.18 (1.42)
English reading motivation (max = 7)	4.70 (1.42)	4.32 (1.51)	4.51 (1.48)
Test importance (max = 7)	5.11 (1.06)	5.07 (1.01)	5.09 (1.03)
Performance compared to peers (max = 7)	4.13 (1.07)	4.32 (0.72)	4.22 (0.82)

Note: The test statistics can be found at https://osf.io/2twzd/.

Table 3. *Reliability and correlations of the proficiency and IQ measures. On the diagonal (in italic) is the cronbach's alpha of each test. All numbers above that are original Pearson correlations, under the diagonal are the correlations corrected for reliability ($r_{xy}/\sqrt{(r_{xx}.r_{yy})}$).*

Tests	**Dutch LexTALE**	**Dutch voc. MC**	**Dutch spelling**	**Eng. LexTALE**	**Eng. spelling**	**MINT**	**QPT**	**Author recogn.**	**Raven**
Dutch LexTALE	*0.63*	0.19	0.33	0.34	0.2	0.21	0.19	0.14	0.04
Dutch voc. MC	0.29	*0.66*	0.25	0.27	0.30	0.30	0.36	0.34	0.18
Dutch spelling	0.46	0.33	*0.87*	0.28	0.61	0.26	0.32	0.25	0.14
Eng. LexTALE	0.44	0.35	0.32	*0.90*	0.54	0.59	0.57	0.28	0.21
Eng. spelling	0.26	0.38	0.68	0.59	*0.93*	0.56	0.55	0.27	0.15
MINT	0.18	0.38	0.29	0.64	0.60	*0.93*	0.63	0.16	0.20
QPT	0.29	0.49	0.38	0.66	0.63	0.72	*0.83*	0.24	0.14
Author recogn.	0.05	0.42	0.27	0.30	0.28	0.17	0.27	*0.97*	0.06
Raven	0.35	0.33	0.22	0.33	0.23	0.31	0.24	0.09	*0.46*

Table 4. *Means, standard deviations and ranges of the scores in the true/false judgement test and the recall test as function of the language in which the text was studied and the test taken.*

	Mean	**SD**	**Min**	**Max**
True/false judgement				
L1 (N = 100)	80.9*	11.8	46.7	96.7
L2 (N = 95)	80.1*	8.7	53.3	96.7
Free recall				
L1 (N = 100)	56.3	14.2	23.3	90.0
L2 (N = 95)	44.1	14.3	15.2	80.0

* Note that chance level for a true/false test is equal to 50%. So the average scores on this test could be compared to a 60% score for the free recall test.

[1] All test statistics (Welch two-sample t-tests and Wilcoxon rank sum tests) for the group comparison can be found at https://osf.io/2twzd/ .

[2] In the self-ratings participants indicated they had more prior knowledge about The Sun (M = 3.3 on a 7-point rating scale) than about Sea Otters (M = 1.8) but that the text about the Sun was experienced as more difficult than the text about sea otters (3.6 vs. 3.1). Both texts were judged to be matched in terms of structural difficulty ($M_{The\ Sun}$ = 2.89 and $M_{Sea\ Otters}$ = 3.03) and in terms of power to interest ($M_{The\ Sun}$ = 4.6 and $M_{Sea\ Otters}$ = 4.5).

[3] There were minor differences between the texts. Recognition test performance was 5.8% better in L1 than in L2 for *The Sun*, but 4.3% worse for *Sea Otters.*

[4] The authors thank an anonymous reviewer for pointing to this issue.

[5] We have since replicated this effect and extended it to intervals of 1 week and 1 month. So, the equivalent performance on the recognition test in L1 and L2 is unlikely to be due to the short time period between the study phase and the test phase.