Studying texts in a non-native language:

A further investigation of factors involved in the L2 recall cost

Heleen Vander Beken, Ellen De Bruyne, Marc Brysbaert

Ghent University, Belgium

Published in the Quarterly Journal of Experimental Psychology

Keywords

Bilingualism, learning from text, English as a medium of instruction, memory

Corresponding Author:

Marc Brysbaeert, Department of Experimental Psychology, Ghent University, Henri
Dunantlaan 2, B-9000 Gent, Belgium
Email: marc.brysbaert @ugent.be

Abstract

With academic internationalisation at full speed, English is increasingly used as a medium of instruction in higher education. The question arises whether unbalanced bilinguals remember study materials in a non-native language (L2) as well as in a first language (L1). In previous studies, we found a disadvantage for students recalling short, expository texts in L2 compared to L1, but no such disadvantage for a true/false recognition test, not even on delayed tests after a month. Since no additional forgetting occurs, the quality of the memory trace seems to be equally strong in both languages and the recall cost might be caused by a lack of production skill in L2. To test this hypothesis, we ran experiments in L1-L1, L2-L1 and L2-L2 conditions with free and cued recall (short open questions). We replicate the L2 free recall cost reported earlier and show that it is due to the encoding in L2 rather than to an L2 production cost. In contrast, we found no significant difference in a new pair of texts with short, cued recall questions, though there was a trend in the expected direction. A summary of the effect sizes obtained so far shows a considerable variety (with rather big confidence intervals), suggesting that the cost of studying in L2 depends on several factors such as study time, test requirement, and language proficiency level.

In a world where modern technology and knowledge are accessible to a high number of people, mutual intelligibility is becoming increasingly important. Hence, more and more people understand and use English as a lingua franca (TNS Opinion & Social, 2012). A similar evolution is happening at European universities, where internationalisation leads to a rising number of exchange students, large numbers of English-written research output and, as a consequence, the use of EMI: English as a medium of instruction (KNAW, 2017; Wächter & Maiworm, 2014). Though the use of EMI has some obvious gains (economically and intellectually), the use of a foreign language can be challenging for teachers and students. The issue is debated in many European countries at the start of every academic year, and in addition the topic is emerging in the scientific literature, for instance by addressing the effects of EMI on the performance and knowledge of students.

A recall cost for studying English L2 texts that is not due to the strength of individual memory traces

Besides listening to lectures in English, an important consequence of the increasing use of English is that students have to read and study text materials in English, even for courses in their native language. In two previous studies, we compared learning in the first language (L1 – in this case Dutch) and a second language (L2 –English) in first-year university students. In the first study (Vander Beken & Brysbaert, 2018) we focused on recognition memory versus recall for short expository biology texts of about 300 words. Undergraduates studied the texts in either Dutch or English and received a test in the same language afterwards. We observed a large *recall cost* (d = .8) for an essay-like test in L2, when students were asked to write down as much as they remembered from the text. At the same time, no L2-cost was found in a recognition task, which consisted of true/false judgements. Note that we used a time limit (7 minutes for the reading phase and another 7 minutes for the testing phase), so that some of the recall cost could be due to a shortage of writing time. An L2 recall cost was also reported by Connor (1984) who observed that significantly more subordinate propositions were recalled by native speakers of English than by L2-speakers. She found no difference for higher-level ideas.

In a second study (Vander Beken, Woumans, & Brysbaert, 2018) we examined one possible origin of the recall cost: the strength of the memory trace in L2 versus L1. Free recall involves an active process of retrieval for which no external cues are given. In a yes/no recognition task, strong memory cues are available, which may help to retrieve memory traces. This type of knowledge has been called *marginal knowledge* (Berger, Hall, & Bahrick, 1999; Cantor, Eslick, Marsh, Bjork, & Bjork, 2014). Perhaps studying in L2 leads to unstable memory traces, in which individual elements can be retrieved when cued, but not recalled without help? Such possibility can be understood within the levels-of-processing theory (Craik & Lockhart, 1972). According to this theory, deeper processing results in better recall. Applied to L1 vs. L2 studying, it implies that deeper processing occurs in L1, yielding an advantage on the recall task, while superficial processing in L2 suffices for the recognition task (Francis & Gutiérrez, 2012).

The reasoning behind the second study (Vander Beken, Woumans, & Brysbaert, 2018) was that stronger memory traces as a result of deeper processing predict longer retention. In other words: If the recall cost in L2 is due to impaired encoding, we expected a steeper forgetting curve in L2 than in L1, because the knowledge decays at a higher rate. This would be especially true for the proportion of knowledge recognised but not recalled (i.e., the marginal knowledge). Therefore, the best choice for investigation was the yes/no recognition task. An additional advantage of this task was that performance in L1 and L2 is the same immediately after studying the text, making it easier to understand differences in the forgetting curve.

Contrary to our predictions, forgetting in L1 and L2 after a day, a week, or a month was very similar. This means that no evidence was found for additional forgetting in a second language compared to a first language. It also meant that the recall cost was not caused by unstable memory traces which could not be recalled but could be recognised. Something else was going on.

Production in L2 as the source of the L2 recall cost?

A second interpretation of the recall cost is that it relates to the requirement of text production. In free recall, participants not only have to retrieve information but also to produce it in a coherent way (even when language mistakes are not taken into account). Yes/no recognition does not involve such a production element. Hence, difficulties in L2 production (writing) could be the origin of the discrepancy between recall and recognition.

There are several reasons why writing skills could affect the L2 recall cost. As Bergsleithner (2010) argued, writing is a process consisting of several subprocesses, such as planning or conceptualising, the actual writing, and revising. Writing is a skill that is acquired at a later age than listening comprehension and that requires active mastery of several aspects of a language, such as lexical knowledge, spelling, and grammar. The fluency with which this information can be retrieved is an important factor in the writing outcome (Schoonen et al., 2002). Also, writing requires metacognitive knowledge, such as text structure and writing strategies (Schoonen et al., 2002; Whalen & Ménard, 1995).

One would expect the complex process of writing to be more challenging in L2. Firstly, L2 proficiency is lower than L1 proficiency. Less knowledge of linguistic aspects (e.g. grammar) affects the comprehension of text but even more so the creation of text. For example, with limited grammar knowledge, it is difficult to construct sentences. Note that for this reason, language mistakes are not punished (as long as they do not obscure meaning) in our research, just like they are not punished in international reading comprehension assessments (e.g. PISA).

Secondly, the level of L2 proficiency also affects the time needed for 'formulation', a sub-process in the writing of an essay-type text (De Larios, Marín, & Murphy, 2001).

A third factor is that L1 is likely to interfere in the L2 writing process, as can be concluded from the observation that L2-writers make errors typical for their L1-background (e.g. Watcharapunyawong & Usaha, 2013).

Fourthly, working memory capacity may be more of an issue in L2. Research has shown that working memory plays an important role in the quality of L2 writing (Bergsleithner, 2010).

Finally, motivation may be less in a non-native language (as found for reading in Vander Beken, 2018) and L2 writing difficulties could lead to writing anxiety (Mat Daud, Mat Daud, & Abu Kassim, 2005), which is known to hinder performance (e.g. in comprehension, Sellers, 2000).

On a more positive note, van Weijen et al. (2009) argued that, while L2 proficiency is related to the quality of short written argumentative essays, it is not related to conceptual activities, by which they mean planning and generating ideas. The authors assume that these activities take place in a language-independent manner. Also Schweppe, Barth, Ketzger-Nöltge & Rummer (2015) argued that although proficient L2 speakers may be worse at verbatim recall, they are equally good at capturing the information conveyed.

Is the L2 recall cost reduced in an L1 test?

If the L2 recall cost is caused by L2 production, then we may expect it to be less severe in L1 production, unless the language switch between encoding and retrieval affects performance by hampering memory retrieval. Such a cost may be predicted on the basis of the encoding-retrieval specificity principle (Tulving & Thomson, 1973), which says that memory benefits from context-congruent conditions at encoding and retrieval. Experimental evidence for this principle in the language domain has been reported in several modalities, such as listening comprehension, word list recall, and episodic memory (Marian & Fausey, 2006; Marian & Neisser, 2000; Matsumoto & Stanny, 2006; Watkins & Peynircioglu, 1983).

If the encoding-retrieval specificity principle holds for memory of texts, a switch from L2 to L1 could be disadvantageous instead of beneficial. On the other hand, if memory is language-independent and stored at a separate, abstract level, as assumed by the dominant theoretical models in bilingualism research (Alba & Hasher, 1983; Dijkstra & van Heuven, 2002; Schank, 1972, 1980), disadvantages because of a language switch may not arise.

Because of the encoding-retrieval specificity principle, researchers of memory in L2 have mainly used L2 tests to keep the language at test (retrieval) congruent to the language of study (encoding). Some authors, however, used L1 as a language for all memory tests.

Four studies addressed the issue by using free recall tests in L1 and asking participants to study texts either in L1 or in L2 (Donin, Graves, & Goyette, 2004; Gablasova, 2014; Horiba & Fukaya, 2015; Roussel, Joulia, Tricot, & Sweller, 2017). They all found better performance in the congruent L1-L1 condition than in the incongruent L2-L1 condition (L1-L1 stands for studying in L1 and testing in L1; L2-L1 for studying in L2 and testing in L1). However, this finding could easily be due to better text understanding in L1 than in L2 at the time of encoding.

A more interesting study was published by Lee (1986), who asked English-Spanish bilingual students to read an easy text in Spanish (L2) and to recall it in Spanish (L2) or in English (L1). Lee reported *worse* performance in L2-L2 than in L2-L1, against the prediction made by the encoding-retrieval specificity principle and in line with the hypothesis that the L2 recall cost may be reduced for recall in L1. A possible reason why Lee's participants performed better in L1 production than in L2 production is that the L2 proficiency of the Spanish-English bilinguals was rather low (first four semesters of Spanish classes). As a result, the participants read a text that was very easy for them (corresponding to a reading level attained by native speakers in the second half of the fifth grade) and for which they had considerable background knowledge in L1.

The specifics of Lee (1986) may be important, because the advantage of L2-L1 production over L2-L2 production was not replicated in another study. Gablasova (2014; see also Gablasova, 2015) asked high school students to learn new, technical words from a text in L2 or in L1. The study happened within the context of CLIL (Content and Language Integrated Learning). CLIL is a high school system in which the students are taught several courses in L2. Half of the target words were tested in L1 and half in L2. Participants remembered fewer words (and in less detail) when they had studied them in L2 than in L1 (even though performance was still reasonably good in L2). In addition, the forgetting rate was higher in L2 on a delayed recall test after one week. Gablasova (2014, 2015) described how "lexical gaps" seem to lead to the omission of certain meaning components: when a word used in a definition of a to-be-learnt word is unknown, it is not recalled. Importantly, participants did not show an effect of language at testing; at least, no effect was reported, arguably because it was not significant. Further specific to the Gablasova study was that most of the technical words were cognates (i.e., had similar forms in L1 and L2).

All in all, there is consistent evidence for better performance in L1-L1 recall than L2-L1 recall, but inconsistent evidence for a difference between L2-L2 and L2-L1. This suggests that the L2 recall cost is due more to encoding than to production. However, the information is suboptimal because none of the studies systematically compared L2-L2, L2-L1 and L1-L1 performance with the same materials and the same participants. The only study we could find that did so was Chen and Donin (1997). A group of 36 Chinese-English bilinguals read short biology texts sentence after sentence in L1 or L2. At four times during the reading they were interrupted to tell what they had just read. At the end of the text, they were asked to recall as much as possible of the entire text. For the text they read in Chinese (L1) the participants were asked to recall in Chinese. For one text read in English (L2), they were asked to recall in English (L2). For the other text read in L2, they were asked to recall in Chinese (L1). Surprisingly, the quality of recall did not differ between the language conditions, although there was a trend towards L1-L1 > L2-L1 > L2-L2. Part of the reason for the lack of significance may be the low power of the study, with a total of only 36 participants (further divided in groups based on study background and proficiency level). In addition, since recall was oral, the L2 production cost may have been less than in written recall. As put by Kormos (2012, p. 390): "Producing 100 words orally might take about a minute in an L2, whereas writing a composition of 100 words might take 30 minutes").

Comparing L1-L1, L2-L2, and L1-L2

To further address the contribution of written text production to the L2 recall cost, we ran two new experiments in which we compared L1-L1, L2-L1, and L2-L2 conditions using validated stimulus materials from previous studies. This allowed us to address three questions:

1. Can the L2 recall cost reported by Vander Beken and Brysbaert (2018) be replicated? This is addressed by comparing conditions L1-L1 and L2-L2.
2. How much of the L2 recall cost is due to being tested in L2? This is addressed by comparing conditions L2-L2 and L2-L1.
3. To what extent are the findings stimulus-specific? To what extent does the recall cost generalize to other materials and test types?

In both experiments we made sure that the studies were properly powered, so that we could draw sensible conclusions. After the empirical part, we also report a meta-analytic summary of the evidence gathered, so that researchers can easily build on it.

Experiment 1

In this experiment we compared a L1-L1 condition (studying in L1and testing in L1), a L2-L1 condition (studying in L2 and testing in L1), and a L2-L2 condition (studying and testing in

L2). The L1-L2 condition was not tested, because we very much expected this condition to be worst (it is hard to write about a new topic in L2 if it has been learned in L1). Such an uninformative finding did not justify the time and resources needed to test a large enough sample.

We used the stimulus materials of Vander Beken & Brysbaert (2018) and Vander Beken et al. (2018). The advantage of using these stimulus materials is that the results can be compared directly. The disadvantage is that all evidence is based on a limited set of materials. The latter aspect will be addressed in Experiment 2.

Method

Design

Since Vander Beken and Brysbaert (2018) and Vander Beken et al. (2018) worked with only two texts, we had to split our study in two smaller experiments if we wanted to use a repeated-measure design (needed to keep the number of participants feasible; a properly powered between-groups experiment with three levels must include at least 435 participants; Brysbaert, 2019). In Experiment 1a we compared the conditions L1-L1 and L2-L1; in Experiment 1b we compared the conditions L2-L1 and L2-L2. Notice that the condition L2-L1 was presented in both subexperiments. This allowed us to make sure that the participants were comparable in both experiments.

Participants

Participants were recruited from Ghent University. First-year psychology students could participate in partial fulfilment of course requirements (about 1/3), other students received payment (about 2/3). Students from language studies or natural science studies were excluded from participation to avoid prior knowledge or high L2 proficiency levels. We further required that all participants were L1-speakers of Dutch (defined here as the dominant language) and had knowledge of English. Since English courses are obligatory in the secondary school system, all participants had studied English for at least four years. In addition, they are regularly exposed to English on (subtitled) television and social media (see De Wilde, Brysbaert, & Eyckmans, 2020 for the importance of out-of-school learning in the acquisition of English as L2). In some of the university courses English handbooks are used as well, even though the teaching happens in Dutch.

We designed the experiment so that we had 80% chance of detecting a main effect of d = .4, which is the typical effect size found in psychology and the smallest effect size with practical implications at the level of the individual (Brysbaert, 2019; Ferguson, 2009). In a repeated-measures design, this requires a minimum of 52 participants, at least when each participant reads only one text per condition (Brysbaert & Stevens, 2018).

In Experiment 1a, a group of 62 students was tested. After exclusion of students with reading problems, students who received a faulty language condition, who reported a different dominant language despite the selection criteria or who were natural science students, 56 valid participants remained. In Experiment 1b, 60 students were tested. After exclusion of a student who had participated in a previous study (despite the selection criteria), 59 students remained.[i]

To check the similarity of the participants in Experiments 1a and 1b we collected language proficiency and working memory data for all participants (see Table 1 and 2). Working memory capacity was measured with the automated operation span task (Unsworth, Heitz, Schrock, & Engle, 2005), administered in E-Prime 2.0.10. The participants' language background and self-rated proficiency was assessed with a selection of questions from the 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). Participants also completed a questionnaire asking about their general motivation and attitude towards testing and reading in L1 and L2, using 7-point Likert scales. For more information regarding this questionnaire, we refer to Vander Beken (2018) or Vander Beken, Woumans & Brysbaert, (2018).

Objective L1 proficiency was measured with a semantic receptive vocabulary test in a multiple choice format with four answer alternatives (listed in Vander Beken, et al., 2018).

[i] For data transparency, we want to add that an additional 26 participants were tested. When testing the 54th participant, a programming error was discovered in the true/false judgment test in one condition (English test for Sea otters): the time limit was set at 240 instead of 420 seconds. Hence, all data in this condition up to that point were discarded and replaced by 26 new participants before any analysis began. Only the interrater reliability is based partially on these data, which should not be problematic, since it is merely a control measure for the objectivity of ratings.

Objective L2 proficiency was measured with the English LexTALE test of receptive vocabulary knowledge for advanced learners of English (Lemhöfer & Broersma, 2012) and the first part (40 items) of the Oxford Quick Placement Test (QPT; 2001), which is considered a measure of general proficiency consisting of multiple choice items of vocabulary and sentence comprehension and grammar.

The results indicated that the L2 proficiency was comparable in both subexperiments (mean score for the LexTALE was 75.6 in Experiment 1a and 74.1 in Experiment 1b) and similar to our previous studies. English and Dutch vocabulary scores and working memory capacity are also similar to those studies. The English LexTALE scores correspond to the B2-level, the upper intermediate proficiency level as described in the Common European Framework (see Lemhöfer and Broersma, 2012, for the correspondence between LexTALE and CEF).

Table 1. Mean scores on the various proficiency and intelligence tests of the participants in Experiments 1 and 2 (standard deviations between brackets).

Variable	Exp. 1a (N = 56)	Exp. 1b (N = 59)
Gender	39F/17M	46F/13M
Age	21.7 (3.21)	20.4 (2.23)
English LexTALE (max = 100)	75.6 (11.82)	74.1 (10.66)
Oxford Quick Placement test (max = 40)	33.0 (3.81)	32.8 (4.04)
Dutch vocabulary MC (max = 75)	49.0 (8.47)	47.3 (6.77)
Operation Span (WM) (max = 75)	60.4 (10.96)	59.7 (10.52)

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

Self-ratings	Experiment 1a	Experiment 1b
General motivation		
Test importance (7)	5.5 (1.22)	5.4 (0.99)
Performance vs. peers (7)	4.2 (0.76)	4.0 (0.81)

Dutch academic reading		
Attitude (7)*	5.2 (1.12)	5.2 (0.83)
Intrinsic motivation (7)*	4.9 (0.99)	4.9 (0.8)
Total motivation (7)*	5.0 (0.92)	5.1 (0.65)
English academic reading		
Attitude (7)*	5.6 (1.04)	4.7 (0.94)
Intrinsic motivation (7)*	4.8 (0.99)	4.5 (0.89)
Total motivation (7)*	4.7 (1.02)	4.6 (0.79)
Opinion about use of EMI (7)*	6.1 (0.97)	6.0 (0.79)
Dutch language skill		
Reading (10)	9.5 (0.69)	9.3 (0.86)
Writing (10)	9.1 (0.97)	8.6 (1.16)
Proficiency (10)*	9.4 (0.65)	9.1 (0.74)
English language skill		
Reading (10)	8.0 (1.21)	7.8 (1.16)
Writing (10)	7.1 (1.28)	8.6 (1.50)
Proficiency (10)*	7.7 (0.87)	7.4 (1.04)

Note: Asterisks indicate sum scores. Likert-scale is indicated between brackets.

Materials

Texts

We used the materials from a previous study (Vander Beken & Brysbaert, 2018), which are adapted versions of 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 (258 words long in English and 248 words in Dutch) and the Sea Otters (279 words in English and 274 words in Dutch). The translations were matched on word frequency in each language. The texts were presented on paper in Times New Roman 10. Line spacing was 1.5 and the first line of every paragraph was indented.

Vocabulary support

Different from Vander Beken and Brysbaert (2018) and Vander Beken et al. (2018), we decided to add explicit vocabulary support. There were two reasons for this. First, students often look up unknown L2 words while studying. Second, we did not want to handicap students in the L2-L1 condition (where explicit translation is needed) relative to the L1-L1 condition. A number of words was selected for each text based on (a) a vocabulary test with words from the Sea otters text that had been administered after the experiment of Vander Beken et al. (2018), and (b) the word frequencies. English words that were not known by at least 80% of the participants in the post-test of Vander Beken et al. (2018) and low-frequency words with a Zipf-value below 3.6 were put on a list together with their L1 translation and given to the students while they were studying the text (for the text on The Sun, there was no post-test, so that only the frequency criterion was used). The list of translations was used for the L2-L1 and L2-L2 conditions. Because the word list might be used by the participants as a memory organizer, a similar list was made for the L1-L1 condition. Here, the L1 words were presented together with a synonym or a hypernym (category label), even though the students were very likely to know the L1 words. For the text about the Sun, the word list contained 10 items; for the Sea otters, it contained 20 items.

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 judgment test. The tests were again taken from Vander Beken and Brysbaert (2018). In the free recall test, participants received the following instruction: (in the language in which they were asked to answer):

“Write down as much as you can remember from the text you have just read. You do not need to copy the text literally (word per word), but give as much information as you can.”

In this way, participants were not asked to literally reproduce the text, but to produce the ideas and to add details where possible.

Roediger and Karpicke (2006) divided their texts into 30 ideas or propositions that had to be reproduced. This list was used as a scoring form for the free recall tests. It was also used to create 46 true/false questions (see Vander Beken & Brysbaert, 2018, for more details). For the true/false 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 given prior to the test and on the top of the screen for every test in the language of the test. Answer options were “True”, “False”, or “I don’t know”. The opt-out option was added in a long-term recognition study (Vander Beken et al., 2018) to avoid guessing. All tests were administered on LimeSurvey, an Open Source web application available through the university. The texts and the tests can be obtained from the authors for research purposes (see also Vander Beken, 2018).

Scoring

In our marking of the free recall tests we followed the correction key laid out by Roediger and Karpicke (2006) and adapted in the previous study (Vander Beken & Brysbaert, 2018).[ii] All memory tests were scored by the first author, using that key. A Dutch-English teacher with test rating experience judged half of the recall protocols (60) in both experiments to control for the reliability of the ratings. Spelling and grammatical mistakes were not punished unless they obscured meaning. The raters were not aware of the language of text, only of the language of test, to avoid any bias. The answers were divided in five categories. A correct answer was scored 1, an incorrect or incomplete answer was scored 0. Because of the cross-lingual conditions, three further distinctions were made: a correct English answer on a Dutch recall form, an incorrect answer that could be traced back to English (e.g. when someone mistranslates an idea), and misinterpretations that must have occurred in the target language.

In order to determine interrater reliability for nominal data rated by two raters, Gwet’s AC1 was calculated (Gwet, 2008). This is robust in cases of high agreement or prevalence (Wongpakaran, Wongpakaran, Wedding, & Gwet, 2013). In our dataset, certain categories

[ii] We also used a new calculation based on an adapted recall key (which split up some ideas in smaller propositions). Although the new key is more refined, it did not affect any of the conclusions. We present the old scores here, because they allow better comparison with our previous studies. More details can be found in Vander Beken (2018).

(correct answers and incorrect/missing answers) made up most of the ratings. Their prevalence could affect the reliability score (Hallgren, 2012) of the commonly used Cohen's kappa. Gwet's AC1 is a value between 0 and 1 and can be interpreted using the same benchmarks as Cohen's kappa. In Experiment 1a, AC1 = 0.88 (SE = 0.009) and in Experiment1b AC1 = 0.89 (SE = 0.008). Because of this very high agreement (> .80 is considered as 'very good' or 'almost perfect' agreement, see Wongpakaran et al., 2013), further analysis was based on the ratings of the first rater.

For the purpose of analysis, the initial categories were transformed to a dichotomous score for correctness. A correct answer in the wrong language also counted as a point, misinterpretations did not. One exception was made for incorrect answers in The Sun. Three ideas in that text contain the numeric term "billion" (10E9), which is an interlingual homograph: in Dutch it translates to "miljard", while the Dutch word "biljoen" means a trillion (10E12). Many participants recalled this idea with the term "biljoen" and were probably unaware of this difference. Since they did recall the contents of the idea, we decided to score these instances as correct. The average of the dichotomous scores was then taken over all propositions, expressed in percentages.

Procedure

All participants were randomly assigned to one of four conditions in which the text order and the language order were counterbalanced, to make sure that the results were not confounded by any of the control variables.

Tests were administered in groups of 6 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 and take two tests (a general recall task followed by detailed questions), with a time limit of 7 minutes for every part (see Vander Beken, 2018, for method and results on the detailed questions). In Experiment 1a, participants were informed that one text would be in English and one in Dutch, but that all tests were in Dutch. In Experiment 1b, they were informed that they would study English texts, but that the test would be either in Dutch or in English. The language of the test was indicated on the page that was visible before studying the text and in the corner of the page with the study text as well.

Texts and recall tests were presented on paper; the true/false test was administered in LimeSurvey.

The experiments started with the studying and testing part. Participants first studied text 1 and after a sort interval of a few seconds (approximately 5 seconds) did a recall test followed by a yes/no recognition test. As the tests were memory tests, participants had to hand in the studied text before they started the tests (contrary to what is sometimes done in reading comprehension tests). The procedure was then repeated for text 2.

Afterwards, participants were given spelling tests (see Vander Beken, 2018 for details on these tests, which are still in development) and the operation span task. They finished by filling in the various questionnaires and vocabulary tests at their own pace. The experiments took approximately two hours in total, plus or minus 15 minutes due to individual or group differences in speed.

Memory performance on the free recall tests

All data is available at https://osf.io/p5b3y/. The scores were computed as in Vander Beken & Brysbaert (2018)[2].

Experiment 1a:

Based on the percentages correct, the mean recall score was 62.5% (SD = 12.89) in the L1-L1 and 54.0% (SD = 14.30) in the L2-L1 condition. Since the data were normally distributed, a one-sided paired t-test was used for analysis (because we expect a higher score in the L1 text condition compared to L2). Performance was significantly better in the L1-L1 condition (t(55) = 4.58, p < .001). A post-hoc calculation of effect size resulted in a Cohen’s d estimate of d = .61, which is considered a “medium” effect, with a confidence interval from 0.23 to 1.00. The Bayes factor B_{10} was 746, which denotes extreme evidence for the alternative hypothesis.

Experiment 1b:

Based on the percentages correct, the mean recall score was 49.9% (SD = 14.41) in the L2-L1 and 51.6% (SD = 13.21) in the L2-L2 condition. Since the means are not in the direction we expected (with a higher score in the L1 recall condition) and the data were normally distributed, a two-sided paired t-test was used. No significant difference was found (t(58) = -0.9, p = .81). A post-hoc calculation of effect size resulted in a Cohen's d estimate of d = -.12, which is considered a "negligible" effect (in the other direction than expected), with a confidence interval ranging from -0.48 to 0.25. The Bayes factor B_{10} was 0.21, which denotes moderate evidence for a null effect.

Comparison between experiments and to previous results:

By juxtaposing the scores in the cross-lingual conditions in Experiment 1a and 1b, we can conclude that both experimental groups performed at a comparable level, though the first group performed slightly better (54.0% vs. 49.9%, mean difference of about one idea). As can be seen in Tables 1-3, this group also scored slightly higher on the English proficiency tests.

The mean recall score in the L1-L1 condition was 62.5% versus 51.6% in the L2-L2 condition. These scores are slightly higher than Vander Beken and Brysbaert (2018; with scores of respectively 56.3 and 44.0) but with a similar difference. It must be remembered that in the present experiments we informed the participants about the language of the upcoming test and gave them vocabulary support.

Memory performance on the true/false tests

Note that these tests were given after the free recall test. Hence, the scores might be inflated by the testing effect or lower because of the longer retention interval.

Experiment 1a

Based on the percentages correct, the mean recognition score was 69.2% (SD = 11.15) in the L1-L1 condition and 65.9% (SD = 12.56) in the L2-L1 condition. Since the data were normally distributed, a two-sided paired t-test was used for analysis. The difference was significant in the t-test (t(55) = 2.14, p = .04) but not when calculated on the basis of Cohen's d (d = .29, confidence interval between -.09 and .67) or according to a Bayesian test (Bayes factor B_{10} = 1.2). Since both the confidence interval and the Bayes' factor do not point towards a true difference and the t-test is borderline significant, this can be considered a very small or non-existent effect.

Experiment 1b

Mean recognition was 66.8% (SD = 13.82) in the L2-L1 condition and 68.2% (SD = 13.50) in the L2-L2 condition. Since the data were normally distributed, a two-sided paired t-test was used for analysis. No significant difference was found (t(50) = -1.93, p = .06). The Bayes factor B_{10} was 0.84.

Comparison between experiments and to previous results:

By juxtaposing the scores in the L2-L1 conditions in Experiment 1a and 1b, we can conclude that both experimental groups performed at a similar level (65.9% vs. 66.8%). The score of the L1-L1 and L2-L2 conditions in Experiments 1a and 1b can be compared to those of Vander Beken and Brysbaert (2018) to check whether the results are replicated (despite being in different groups). The mean score in the L1-L1 condition (experiment 1a) was 69.2% (SD =11.15) and in the L2-L2 condition (experiment 1b) it was 68.2% (SD = 13.50). In the same conditions of Vander Beken and Brysbaert (2018), they were 80.9% (SD = 11.8) and 80.1% (SD = 8.7). Hence, the scores in the present study are lower than those from the previous study, probably due to a smaller guessing factor and to the fact that the recognition test was presented after the recall test instead of immediately after learning. Nonetheless, the L1-L1 and L2-L2 conditions were very close to each other once, replicating the pattern published in Vander Beken and Brysbaert (2018) and Vander Beken et al. (2018). Table 3 summarizes the findings.

Table 3. Overview of the memory scores in percentages.

Memory tests	L1-L1	L2-L1	L2-L1	L2-L2
True/false	69.2	65.9	66.8	68.4
True/false (Vander Beken & Brysbaert, 2018)	80.9			80.1
Free recall	62.5	54.0*	49.9	51.6
Free recall (Vander Beken & Brysbaert, 2018)	56.3			44.0

Note: Asterisks indicate significant differences based on a Bayes test.

Discussion

In Experiment 1 we tested to which extent the free recall cost in L2, reported by Vander Beken and Brysbaert (2018), can be ascribed to difficulties with L2 production or to an impaired mental model as a result of difficulties with L2 comprehension or encoding. The same materials were used.

We replicated the poorer recall performance in L2-L2 than in L1-L1 ($d \approx .6$) and the fact that there was no similar L2 cost in the yes/no recognition test. More importantly, we found that the L2-L1 conditions were much closer to the L2-L2 condition than to the L1-L1 condition (see Table 3). This indicates that the L2 recall cost is not primarily due to L2 production, but is a result of studying a text in L2.

Another interesting observation is that we found no evidence for the encoding-retrieval specificity principle, according to which L2-L2 should have resulted in better performance than L2-L1. So, the L2 test cost seems to outweigh any encoding-retrieval specificity recall benefit that might be present (attentive readers may notice that there is a trend towards an encoding-retrieval specificity benefit in the recognition test). We will come back to the significance of these findings in the general discussion.

A question we first want to address is to what extent the findings are limited to the stimulus materials used. Indeed, all our research so far has made use of two texts taken from Roediger and Karpicke (2006). These texts have the following characteristics:

- They are dense (30 new units of information in 270 words).
- They are on topics students are less familiar with.
- They arguably are on topics students find less interesting than their own studies.

Although such texts are typical for encyclopedias (and summaries compiled by students for their exams), they may not be very characteristic of the textbooks students use. Such textbooks usually are more verbose (using sentences and paragraphs to explain ideas rather than single phrases) and build largely on previous knowledge students acquired. In addition, students may be more motivated for topics typical of their own study degree than for sea otters and the sun. Finally, exams are rarely full recall tests (except maybe for essay-type questions). In the next experiment we investigate to what extent the L2 recall cost is observed under other, more typical study conditions.

Experiment 2

In this experiment we examine to what extent the L2 recall cost is observed in a situation in which students study a section from a textbook and are asked short, open-ended questions.

The same design was used as in Experiment 1. That is, the study was split up into two smaller experiments with one repeated-measure variable. In Experiment 2a, participants studied one text in English and one in Dutch and were tested in Dutch for both texts (i.e., L1-L1 and L2-L1). The language of the texts was counterbalanced over participants, to avoid text-specific effects. In Experiment 2b, participants received both texts in English, but received one test in Dutch and one in English (i.e., L2-L1 and L2-L2). The language of the tests was again counterbalanced over participants. Notice that the L2-L1 condition, as in Experiment 1, was the same in both subexperiments (and hence can be compared across experiments).

Different from Experiment 1 was that participants were not informed beforehand about the language of the test they would get. This is because chronologically Experiment 2 was administered before Experiment 1 (as mentioned in Vander Beken, 2018). We present it here after Experiment 1 because it streamlines the narrative.

Method

Participants

The data in this study were collected in a subgroup of the participants tested in Vander Beken et al. (2018), as this was the only way to collect large enough numbers in a PhD period of 3 years. The data were collected after the experiments discussed in Vander Beken et al. (2018), so that they did not influence the results of that study.

The same participation requirements were used as in Experiments 1. All in all, data of 53 participants were retained for Experiment 2a (from an initial number of 68) and 63 in Experiment 2b (from an initial 78).

As in Experiment 1, we collected proficiency and working memory data for all participants. One difference was that we gave the multiple choice Vocabulary Size Test (Nation & Beglar, 2007; 14,000 version) instead of the Oxford Placement Test. The data can be found in Table 4. The levels of L2 proficiency, as measured with LexTALE (M = 74), are comparable to those in Table 1. The score of 96 on the vocabulary size test means our participants knew 9,600 of the 14,000 most frequent word families. This is a rather high level of proficiency. For advanced degrees, Paul Nation (2006) estimates that non-native students require a receptive vocabulary size of at least 8000 to 9000 word-families. Importantly, the participants were very comparable in Experiments 2a and 2b, simplifying the comparison of the experiments.

Table 4. Mean scores on the various proficiency and intelligence tests of the participants in experiments 2a and 2b (standard deviations between brackets).

Tests	**1**	**2**
Gender	39F/13M (1 NA)	47F/14M (2 NA)
Age	19.1 (1.75)	20.3 (5.06)
Dutch vocabulary MC (max = 75)	46.5 (7.55)	47.9 (7.60)
English LexTALE (max = 100)	73.5 (9.39)	73.9 (10.80)
English Vocabulary Size (max = 140)	95.7 (11.83)	95.9 (13.38)
Operation Span (WM) (max = 75)	57.3 (13.17)	58.5 (11.07)
Self-rating of Dutch lang. skill (max = 10)	9.6 (0.88)	8.9 (0.75)
Self-rating of English lang. skill (max = 10)	7.6 (1.07)	7.1 (1)

Note: Underlined variables indicate sum scores. Data files are available at **https://osf.io/c67ya**.

Materials

Text materials

In our previous studies, we worked with two highly controlled texts matched between languages on content and word frequency measures. For the present study, we wanted the texts to be real educational material for students, to increase the ecological validity of the experiments. Therefore, we decided to use two matched excerpts from psychology books, written by the same author (we eliminated a few differences, such as an entire sentence in one text that was not mentioned in the other text). The books were a Dutch handbook *Psychologie* (Brysbaert, 2016) and an English handbook *Historical and Conceptual Issues in Psychology* (Brysbaert & Rastle, 2009; note that this book was co-authored by a native speaker of English). We made sure that the excerpts were not studied in the students' courses yet, and we tested them early in the academic year to avoid strong prior knowledge.

The first texts excerpt was *The experiments of Zajonc and colleagues on the perception of emotions* (referred to as *Zajonc* from here) covering the topic of subliminal perception and masked priming. The English version was 488 words long, the Dutch 395 words. The second text, called *Myth busting: is unconscious processing dangerous?* (referred to as *Myths* from here) discussed experiments with subliminal messages. The English version was 517 words long, the Dutch 537. All texts (English and Dutch) can be found in the supplemental materials.

Test materials

We formulated three open questions (on 1 or 2 points each) based on the texts, with a maximum total score of 5 points per test. The answer keys were established independently by two authors of the present article, one of whom is the author of the books and a seasoned exam marker, with the purpose of creating robust marking schemes. Each test contained one reproductive question (e.g. "What is semantic priming?") and two questions that required more inferential or applied reasoning such as "Will the same results be found in a different group of participants"? The questions and answer keys are included in the supplemental materials.

Scoring

All memory tests were scored on the basis of the answer keys by the first author and a second rater affiliated to the Department of Experimental Psychology (thus having sufficient knowledge of the topic). Since we were interested in memory retrieval rather than writing skill, we adopted the guideline for PISA tests (see appendix in Cartwright, 2012) not to punish spelling and grammatical mistakes unless they obscured meaning. The raters were not informed about the language of the text studied, to avoid any bias (though it was impossible to avoid some indications of the condition due to Anglicisms or English words in the Dutch recall protocols of the L2-L1 condition). A correct answer was scored 1; a partially correct answer 0.5. The mean interrater reliability over all tests and questions was r = .76 (range per text and condition: .70 to .85). The scores were summed to test scores of maximum 5 points (for which the interrater reliability score was r = .82). The dependent variable was the average of the summed scores of both raters per participant over the questions.

Procedure

Participants were tested in groups of 33 at most. They registered online for an experiment that entailed two lab sessions and some homework filling in questionnaires online. In the first lab session, students studied four texts, starting with two short expository texts and tests about biology topics and their respective recognition memory tests (Vander Beken et al., 2018). After these tests, they received the *Zajonc* text and test, and then the *Myths* materials.

The participants were informed that they had to study a text and would receive a test afterwards, with a 7-minute time limit in both phases. They were allowed to highlight sections of the texts or to make notes, but only on the text itself, which they had to put aside once their study time was up. After the test phase, the procedure (study phase – test phase) was repeated for the second text.

The texts were presented on paper, in Times New Roman, font size 10, and were divided in two or three paragraphs (one page). The tests were administered online using LimeSurvey. There were only a few seconds (approximately 15 seconds) between studying the text and filling in the retrieval test (i.e., the time needed to put aside the text and open the test; instructions were in group). LimeSurvey was also used for the LexTALE and the Dutch vocabulary test. Nation and Beglar's (2007) Vocabulary Size Test was administered via the original website www.vocabularysize.com. Participants either did these tests online during session 1 or at home between sessions 1 and 2.

Results

Memory test scores in Experiment 2a

Based on the average scores, the mean was 3.1 (SD = 1.15) for the L1-L1 test and 2.8 (SD = 1.15) for the L2-L1 test (see Table 5). The data were right-skewed, but since our sample size was sufficiently large (N = 53), a paired samples t-test was used for analysis.[iii] There was no evidence that the scores were higher in the L1-L1 conditions than in the L2-L1 condition (t(52) = 1.59, p = 0.12). A post-hoc calculation of effect size resulted in a Cohen's d estimate of d = .20, which is considered a "small" effect. The confidence interval of the effect size ranged from -0.19 to 0.59. Next to hypothesis testing, we also ran a Bayesian analysis to assess to which extent the observed distribution corresponded to the expected distribution of the null hypothesis. The Bayes factor B_{10} was 0.49, which signals as much evidence for the null hypothesis as for the alternative hypothesis.

Memory test scores in Experiment 2b

Mean recall was 2.8 (SD = 1.2) for L2-L1 and 2.6 (SD = 1.24) for L2-L2 (see Table 5). Since the data were normally distributed, a paired t-test was used for analysis. No significant difference was found (t(62) = 0.83151, p = .41). A post-hoc calculation of effect size resulted in a Cohen's d estimate of d = .12, which is considered as a "negligible" effect. The confidence interval of this effect size ranged from -0.24 to 0.47. The Bayes factor B_{10} was 0.19 here, which can be interpreted as moderate evidence for the null hypothesis.

Table 5. Means and standard deviations (between brackets) of the memory scores in experiment 1 (correct points out of 5).

	Experiment 1: all tests in L1		**Experiment 2: all texts in L2**
L1-L1	3.07 (1.15)		
L2-L1	2.77 (1.15)	**L2-L1**	2.82 (1.20)
		L2-L2	2.63 (1.24)

Discussion

[iii] To make sure, we also ran a non-parametric test. The Wilcoxon signed-rank test did not yield different results from the T-test.

In experiments 2a and 2b, we examined whether participants would perform better on an L1 recall test of a text studied in L2 than on an L2 test for the same materials. Different from Experiments 1a and 1b was that (1) short recall questions were asked instead of one large essay-type question, (2) the texts were less information dense, and (3) the topics were closer to the students' interests and existing knowledge.

Contrary to the results reported in Experiments 1a and 1b and in Vander Beken and Brysbaert (2018), this time no significant L2 language cost was found, even though we had an a priori power of 80% to detect effect sizes as small as d = .4. On the other hand, there was a trend in the direction of L1-L1 > L2-L1 > L2-L2, similar to what was observed by Chen and Donin (1997) and in Experiment 1.

As in Experiment 1, we found no evidence for the encoding-retrieval specificity principle, according to which L2-L2 should have resulted in better performance than L2-L1, again indicating that an L2 production cost outweighs any encoding-retrieval specificity benefit. At the same time, it looks as if every addition of L2 to the task (be it input or output) causes a small drop in performance. Arguably, in combination with the results of Experiment 1, this points towards some difficulties as a result of L2 both at the encoding and the production level.

The diverging findings of Experiments 1 and 2 create an interpretation problem, because it looks like the L2 costs are function of a number of variables, such as the (1) the difficulty of the texts, (2) the type of test used, and arguably (3) the proficiency level of the participants, and (4) the time given to study. These axes make a large problem space, which so far has been sampled very sparsely. In order to bring more clarity to the issue and to make the first step in delineating the distribution of L2 costs, we decided to end with a meta-analysis summarizing what has been found so far (McShane & Böckenholt, 2017). Hopefully, this will help future researchers to explore the problem space more systematically.

A meta-analytic summary of what has been found so far

In Experiment 1 and 2, we calculated the effect sizes, similarly to previous studies (e.g. Vander Beken & Brysbaert, 2018). To improve our understanding of the L2 disadvantages and the conditions in which they appear, a direct comparison between the effect sizes in the present and other available studies would be valuable. Therefore, we have calculated effect sizes of all text recall studies that compared L1 and L2 and that reported enough statistics to derive the effect size.

Method

Because the question involves pairwise comparisons, we can work with Cohen's d (or *Hedges's g* if we want to correct for small sample sizes). A challenge, however, is that the previous designs were between-subjects designs and the present are within-subjects designs. Whereas Cohen's d is straightforward to calculate for between-subjects designs, there are two possible calculations for within-subjects designs, sometimes called d_z and d_{av} (Brysbaert, 2019; Lakens, 2013). The first measure, d_z, is the one traditionally calculated for power analysis, as done in the results sections. It assumes that the overall differences between participants do not matter. The second measure, d_{av}, takes the differences between participants into account, so that the effect size is comparable to the one obtained in between-subjects designs. D_z is equal to d_{av} when the correlation between the measures in the two conditions across participants is $r = .5$. If the correlation is higher, d_z will be larger than d_{av}; if the correlation is lower, d_z will be smaller than d_{av}. Often the correlation is higher than .5, particularly for experiments involving reaction times (because some participants respond much faster than other). This means that for these studies the effect sizes are likely to be bigger in within-subjects designs than in between-subjects designs if based on d_z. On the other hand, the correlation tends to be closer to .5 for memory studies, so that the discrepancy between the effect size measures may be less of an issue here.

There are several ways to calculate d_z and d_{av}. The simplest equations for d_z are:

$$d_z = \frac{M_{diff}}{\sqrt{\frac{SD^2{}_{diff}}{N-1}}}, \text{ or } d_z = \frac{t}{\sqrt{df}}$$

Applied to the data of Experiment 1a, we can calculate that:

$$d_z = \frac{t}{\sqrt{df}} = \frac{4.58}{\sqrt{55}} = .62$$

The equation for d_{av} is:

$$d_{av} = \frac{M_{diff}}{\frac{SD_1+SD_2}{2}}.$$

(If you have access to the raw data, you can simply calculate d_{av} by using algorithms for unrelated samples, like when you calculate d for a between-subjects design).

Applied to Experiment 1a, it gives:

$$d_{av} = \frac{62.5-54.0}{\frac{12.89+14.30}{2}} = .63.$$

Both values are close together, meaning that the correlations between the participant scores in the L1-L1 and L2-L1 conditions of Experiment 1a were about r = .5. The same is true for the other effect sizes reported here:

Experiment 1b:

$$d_z = \frac{t}{\sqrt{df}} = \frac{-.90}{\sqrt{58}} = -.12 \qquad d_{av} = \frac{-2.7}{\frac{14.4+13.2}{2}} = -.20.$$

Experiment 2a:

$$d_z = \frac{t}{\sqrt{df}} = \frac{1.59}{\sqrt{52}} = .22 \qquad d_{av} = \frac{.3}{\frac{1.15+1.15}{2}} = .26.$$

Experiment 2b:

$$d_z = \frac{t}{\sqrt{df}} = \frac{.83}{\sqrt{62}} = .11 \qquad d_{av} = \frac{.2}{\frac{1.20+1.24}{2}} = .16.$$

Results

Table 6 shows the effect sizes we were able to calculate for the various studies that tested free or cued recall of L2 texts in a quantitative manner and that provided enough information for us to derive an effect size. For example, the study by Chen and Donin (1997) only provides significant test statistics and figures with group means split up for conditions (e.g. background knowledge groups) and types of information. Even the group means and SDs are not provided. Hence, it is impossible to provide an estimation of the effect size. Similarly, Connor (1984) does not report standard deviations of her conditions or t-values for her tests.

The Cohen's ds reported in our papers with ANOVAs are based on post-hoc calculations of the relevant contrasts. In the same way, we selected the relevant means for an effect from other research papers. For example, for the paper by Roussel and colleagues (2017), the means of the L2-L1 and L1-L1 condition were taken, while the original analysis also contained a condition with an L2 text and a translation. For detailed test statistics on all language conditions, one can turn to the original papers. For studies by other authors, Cohen's d was calculated by using an online effect size calculator (Lenhard & Lenhard, 2016).

< insert Table 6 about here >

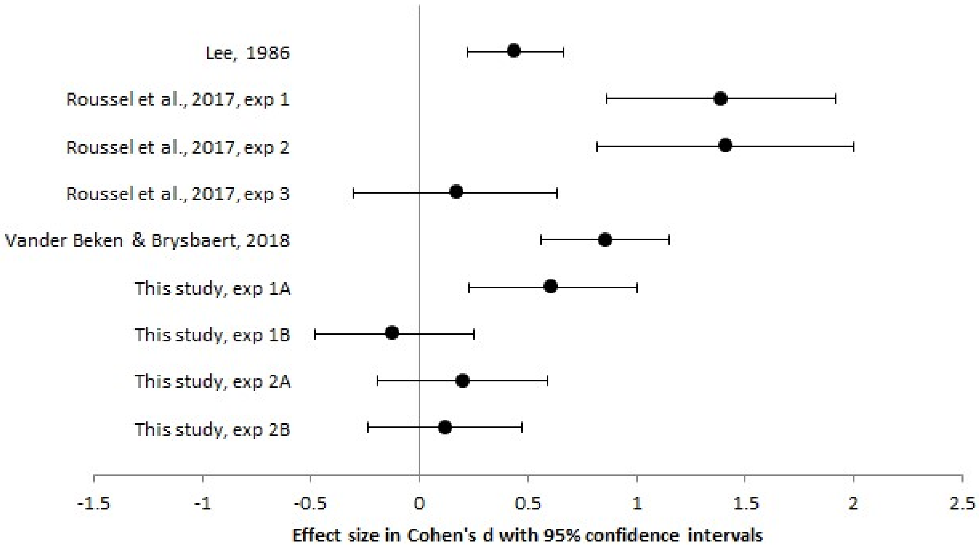

Figure 1: Forest plot of the recall costs obtained so far in comparisons of L1-L1, L2-L1 and L2-L2 conditions. See Table 6 for more details.

Discussion

The most important finding from the meta-analysis is that free recall (Connor, 1984; Lee, 1986; Vander Beken & Brysbaert, 2018; Experiment 1 here) is not the only task in which a clear L2 disadvantage is found. Large effects were also found in Gablosova (2014, 2015) and Roussel et al. (2017). These authors investigated language and content learning within the context of CLIL and tested the participants' content knowledge with questions probing cued recall about specific parts of the text. Gablasova presented texts in L1 or L2 and asked about specific technical words explained in the texts. Roussel et al. (2017) had three conditions (L2 text, L2 text with L1-translation, and L1 text. They tested the participants' content knowledge with L1 questions probing cued recall and observed that performance on the knowledge tests was best in the L1 condition, and worst in the L2 condition without translation support. The effect was significant in the first two studies run, but not in the last experiment involving students of computer sciences. Our calculations show an effect size of Cohen's d = .2 in the last condition with an upper limit of the confidence interval close to the lower limit of the other intervals. So, it is not clear to what extent this study really deviates from the other two. It is reasonable to believe that the participants in the first two studies of Roussel et al. studies performed worse in L2 than our participants, since their proficiency level was lower (B1 vs B2 in the CEFR framework). The computer science students in Experiment 3 were probably more trained in using English materials or terminology.

At the same time, three studies reported in the present manuscript have effect sizes meandering around zero (and we have to take into account the ambiguous status of Experiment 3 of Roussel et al. 2017). A problem with these studies is that we do not know whether they point to no differences or to small effects that could still be of value. Durlak (2009), for instance, warns against overlooking small effect sizes that can be of practical importance. An effect size of d = .2 can be of policy interest if it is related to educational achievement. Another problem for these studies is that we do not know how many other studies with null effects have been run but not reported in the literature, because there are less incentives for authors, reviewers and editors to publish them (Franco, Malhotra, & Simonovits, 2014; Norris, 2015).

General discussion

In this article we sought to answer three questions:

1. Can the L2 recall cost reported by Vander Beken and Brysbaert (2018) be replicated?
2. Is the recall cost reported by Vander Beken and Brysbaert (2018) due to being tested in L2?

3. Does the recall cost generalize to conditions other than the ones examined by Vander Beken and Brysbaert (2018)?

The answer to the first question was a clear yes (see Experiment 1a), as could be expected given that Vander Beken and Brysbaert (2018) found a large effect size with a well-powered study. In addition, we showed that the effect persists when participants are given vocabulary support and are told in advance which language will be used for the test. Participants who have to study in L2 and are given a free recall test in L2 perform substantially worse than participants who are allowed to study in L1 and take a free recall test in L1 (see also Connor, 1984).

Still related to question 1, we also replicated the finding that the L2 cost is not observed in a yes/no recognition task, arguably because yes/no questions provide enough cues to find memory traces that are not accessible otherwise. Interestingly, Gablasova (2014) also mentioned the possibility that some knowledge which her participants could not express, might be accessible in a receptive manner. This passive knowledge cannot be recalled actively but can be recognized when strong memory cues are given. This may explain why gaps in the mental model do not affect true/false questions, but impede recall and possibly also performance on demanding multiple-choice questions. Future research will have to disentangle these possibilities by testing both active and passive knowledge at the word and the text level.

The answer to the second question of our study was no. Participants did not perform better in an L2-L1 condition (with testing in L1) than in an L2-L2 condition (with testing in L2), at least not when language mistakes were accepted if they did not obscure the meaning. Both conditions with L2 learning were worse than the condition with L1 learning. This indicates that the L2 recall cost was primarily due to L2 encoding and not to L2 production. This is in line with Gablasova (2014, 2015) but not with Lee (1986). As we argued in the introduction, we think the deviating finding of Lee may be due to the fact that he tested students on an easy text for which the participants arguably had a much background knowledge in L1.

At the same time, we found little evidence for an encoding-retrieval specificity benefit in free recall: Participants did not perform better in L2-L2 than in L2-L1. There was even a tendency towards the opposite. Only in an immediate yes/no recognition test was there some evidence for an effect due to the similarity of the words in text and test (Table 5).

The most likely explanation of the full set of findings in Experiment 1 is that when we study a text for a subsequent memory test, the information is largely translated into a language-independent, abstract memory code. This code is used for recall (hence the absence of an encoding-retrieval specificity benefit) and for fact retrieval (hence, the language-independent performance on the yes/no recognition tests). However, the code is richer and better organised when texts are studied in L1 than in L2 (also see Gablasova, 2014; Roussel et al., 2017), helping performance in tasks without memory cues. An interesting perspective on this is provided by 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. 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 studies in L1 is richer than that of a text studied in L2.

As for the third question we examined (about the generality of the L2 cost), the answer is more equivocal. Whereas the findings are very coherent for the stimulus materials used in Experiment 1 and the previous experiments we ran, we did not obtain a significant L2 recall cost for the stimulus materials used in Experiment 2. In this experiment both the text and the test were changed, to make the experiment more similar to a typical study-exam situation for our participants. In particular, the texts were made less demanding by making them more verbose and closer to the interests and the knowledge of the participants. Also the tests were made less demanding, because we replaced the essay-type free recall question by three questions with rather short answers. Under these circumstances, we still found some evidence for L1-L1 > L2-L1 > L2-L2, but the differences were small and no longer statistically significant.

The null effect of Experiment 2 suggests that the L2 cost is a gradual effect, depending on several factors, as also suggested by the conflicting findings of Lee (1986) and Gablasova (2014), discussed above. For what it is worth, these are the factors we think are at play.

A first factor likely to be involved in the L2 cost, is the L2 proficiency of the participant relative to the text difficulty. Based on the Landscape model, we can predict that the cost will be particularly high when participants do not (yet) have rich semantic representations for the L2 words they have to digest. As they become more experienced, performance will come closer and closer to L1 performance. Note in this respect that L2 proficiency is often more topic specific than L1 proficiency. Non-English-speaking psychology researchers may feel more at ease reading English texts on psychology-related topics than on, for instance, house-building-related topics.

Chen and Donin (1997) made an interesting remark in this respect. They proposed linguistic distance as an explanation for their results. For bilinguals with a great linguistic distance between their languages, little overlap in lexical and syntactic representations exists and this may make it more difficult to deeply encode the stimulus materials, also because the discrepancy is likely to put extra strain on working memory. So, linguistic distance may well be a second factor influencing the size of the L2 processing cost.

A third factor we have seen is the type of test used. The more the test relies on a rich network of connections between learned facts (like free recall of an entire text), the more outspoken the L2 cost will be. This means that lecturers are likely to underestimate the knowledge of L2 students if all exams consist of essay-type questions.

A fourth factor is the cost of L2 production. Although we found little evidence for a L2 production cost with our proficiency group and system of marking, there are good reasons to expect that performance often will be worse when examinees have to respond in L2 than when they are allowed to answer in L1. This will be particularly the case when participants have considerable background knowledge of the topic in L1 (Lee, 1986).

A final factor involved is the effort invested in the studying. For bilinguals similar to the ones studied in Experiments 1 and 2, a typical finding is that they require 20% more time to read and study the same materials in L2 than in L1 (Cop, Dirix, Drieghe, & Duyck, 2016; Dirix et al., 2020). So, time pressure is likely to hurt L2 performance more than L1 performance. Alternatively, having plenty of time and being well motivated may occasionally turn L2 studying into a "desirably difficulty" (e.g. Metcalfe, 2011), because the more demanding learning conditions may make the memory traces stronger, so that test performance becomes better in L2 than in L1.

Bringing the various pieces together will be one of the main challenges for future research, because each study can only address a few cells of the mosaic. In addition, if we want useful information, each study must be properly powered (requiring a considerable investment of time and energy). Still, we think the enterprise is feasible and we present Table 6 as a summary of what has been achieved so far and as a stepping stone for further mapping, so that psychologists can give advice for a wide range of educational situations.

All in all, our findings of an L2 cost in some situations but not in other indicate that the use of English as a medium of instruction (EMI) need not be a problem, as long as certain factors are taken into account. For example, providing sufficient time is a key element to successful content learning in L2. Obviously, the most challenging situation is one in which a non-English student is embarking on a new topic, taught entirely in English and tested in English with exams capitalizing on free recall (e.g., essay-type of questions). Or as Roussel et al. (2017, p. 70) argued: "we may need to be concerned by what happens in situations where students […] are exposed to academic content in this foreign language without any foreign language instructional support". Although this may be the end goal of EMI, it seems wiser not to throw in students at the deep end, but to prepare them via a series of less demanding, intermediate goals. These include the use of texts that are appropriate for the students' proficiency level, and the use of questions that put less demand on uncued recall and English text production. Something else that is likely to help is to provide students with more explicit information about the mental model they are supposed to build (e.g., by providing an L1 skeleton or by using other forms of advance organizers; Ausubel, 1960). In other words, the decision to use a foreign language for higher education should not be taken lightly, but in relation to the background of the students, their capacities, and what will be expected from them in their future professional environments.

>

Table 6. Overview of experiments in five papers which directly compared recall of texts in the native and a non-native language.

Paper	Language conditions	L1 and L2	Language factor	Type of recall	Level of proficiency	Min./100 words°	N_{group}	d (CI)
Lee, 1986	L2 - L2 L2 - L1	English Spanish	between	Free	1st & 2nd-year FL uni students (no CEFR ref)	Self-paced	160 ’	0.44 (0.22 –0.66)
Gablasova, 2014, 2015	L1 - L1/L2 L2 - L1/L2	Slovak English	between	Free	B2 (CEFR)	1.8’’	32	0.59 (0.09 – 1.09)
Roussel et al., 2017 Experiment 1*	L1 - L1 (L2+trans – L1) L2 - L1	French German	between	Cued	B1 (CEFR)	6.2**	34	1.39 (0.86 – 1.92)
Roussel et al., 2017 Experiment 2*	L1 - L1 (L2+trans - L1) L2 - L1	French English	between	Cued	B1 (CEFR)	6.2	28	1.41 (0.82 – 2.00)
Roussel et al., 2017 Experiment 3*	L1 - L1 (L2+trans – L1) L2 - L1	French English	between	Cued	B1 (CEFR)	6.2	36	0.17 (-0.3 – 0.63)
Vander Beken & Brysbaert, 2018	L1 - L1 L2 - L2	Dutch English	between	Free	72 (LexTALE) = B2 (CEFR)	2.6	97	0.86 (0.56 – 1.15)
This study, experiment 1A	L1 - L1 L2 - L1	Dutch English	within	Free	76 (LexTALE) = B2 (CEFR)	2.6	62	0.61 (0.23 – 1.00)
This study, experiment 1B	L2 - L1 L2 - L2	Dutch English	within	Free	74 (LexTALE) = B2 (CEFR)	2.6	59	-.12 (-0.48 – 0.25)

This study, Experiment 2A	L1 - L1 L2 - L1	Dutch English	within	Cued	73 (LexTALE) = B2 (CEFR)	1.4	53	0.26 (-0.14 – 0.66)
This study, Experiment 2B	L2 - L1 L2 – L2	Dutch English	within	Cued	74 (LexTALE) = B2 (CEFR)	1.4	63	0.12 (-0.24 – 0.47)

* We only take the results on a content post-test into account here. Language and transfer post-tests show different results, but are not our main concern in this overview.

° Length of texts based on English versions, even if there is an experiment which includes another non-native language

** This does not take into account the L1 text in the translation condition (which ought to be about twice the number of words)

’ This study compared several proficiency groups and conditions with and without explicit instructions, resulting in 16 groups. Each group contained 20 participants, which means that the total language groups contained 160 participants. So we consider $N_{language}$ = 160.

'’ Estimated on the basis of 10 min reading plus 5 min listening to the text; only immediate posttest.

Funding statement

This study was supported by a GOA grant from the Research Council of Ghent University (LEMMA Project).

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*(2), 203–231. http://doi.org/10.1037/0033-2909.93.2.203

Ausubel, D. P. (1960). The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of Educational Psychology*, *51*(5), 267–272. http://doi.org/10.1037/h0046669

Berger, S. A., Hall, L. K., & Bahrick, H. P. (1999). Stabilizing access to marginal and submarginal knowledge. *Journal of Experimental Psychology*, *5*(4), 438–447.

Bergsleithner, J. M. (2010). Working memory capacity and L2 writing performance. *Ciências & Cognição*, *15*(2), 2–20.

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. (2019). *Basic statistics for psychologists* (2nd ed.). London: Palgrave.

Brysbaert, M., & Rastle, K. (2009). *Historical and Conceptual Issues in Psychology* (Second). Pearson Education.

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. http://doi.org/10.3758/s13421-014-0462-6

Cartwright, F. (2012). Technical feasibility of Reporting YITS 2010 Skill Assessment Results on the PISA 2000 Reading Scale. *OECD Education Working Papers*, *69*. http://doi.org/http://dx.doi.org/10.1787/5k9fhndspvf1-en

Chang, Y. (2011). The relation between time spent on the written recall task and the memory of L2 text. *Reading and Writing*, *24*, 903–919. http://doi.org/10.1007/s11145-010-9231-5

Chen, Q., & Donin, J. (1997). Discourse Processing of First and Second Language Biology Texts: Effects of Language Proficiency and Domain-Specific Knowledge. *The Modern Language Journal*, *81*(2), 209–227.

Connor, U. (1984). Recall of Text: Differences between First and Second Language Readers. *TESOL Quarterly*, *18*(2), 239. http://doi.org/10.2307/3586692

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. http://doi.org/10.3758/s13428-016-0734-0

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. http://doi.org/10.1016/S0022-5371(72)80001-X

Davis, J. N., Lange, D. L., & Samuels, S. J. (1988). Effects of text structure instruction on foreign language readers' recall of a scientific journal article. *Journal of Reading Behavior*, *20*(3), 203–214. http://doi.org/10.1080/10862968809547639

De Larios, J. R., Marín, J., & Murphy, L. (2001). A Temporal Analysis of Formulation Processes in L1 and L2 Writing. *Language Learning*, *51*(3), 497–538. http://doi.org/10.1111/0023-8333.00163

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition, 23(1)*, 171-185. DOI: https://doi.org/10.1017/S1366728918001062

Dijkstra, T., & van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3). http://doi.org/10.1017/S1366728902003012

Dirix, N. Vander Beken, H., De Bruyne, E., Brysbaert, M., & Duyck, W. (2019). Reading text when studying in a second language: An eye-tracking study. Reading Research Quarterly. Advance publication available at https://ila.onlinelibrary.wiley.com/doi/full/10.1002/rrq.277.

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.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538. http://doi.org/10.1037/a0015808

Francis, W. S., & Gutiérrez, M. (2012). Bilingual recognition memory: Stronger performance but weaker levels-of-processing effects in the less fluent language. *Memory & Cognition*, *40*(3), 496–503. http://doi.org/10.3758/s13421-011-0163-3

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505. http://doi.org/https://doi.org/10.1126/science.1255484

Gablasova, D. (2014). Learning and retaining specialized vocabulary from textbook reading: Comparison of learning outcomes through L1 and L2. *Modern Language Journal*, *98*(4), 976–991. http://doi.org/10.1111/modl.12150

Gablasova, D. (2015). Learning technical words through L1 and L2: Completeness and accuracy of word meanings. *English for Specific Purposes*, *39*, 62–74. http://doi.org/10.1016/j.esp.2015.04.002

Lenhard, W. & Lenhard, A. (2016). *Calculation of Effect Sizes*. Retrieved from: **https://www.psychometrica.de/effect_size.html**.

Horiba, Y., & Fukaya, K. (2015). Reading and learning from L2 text: Effects of reading goal, topic familiarity, and language proficiency. *Reading in a Foreign Language*, *27*(1), 22–46.

Joh, J. (2006). What Happens When L2 Readers Recall ? *Language Research*, *42*(1), 205–238.

KNAW. (2017). *Nederlands en/of Engels, Taalkeuze met beleid in het nederlands hoger onderwijs*. Amsterdam.

Kormos, J. (2012). The Role of Individual Differences in L2 writing. *Journal of Second Language Writing*, *21*(4), 390–403. http://doi.org/https://doi.org/10.1016/j.jslw.2012.09.003

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(NOV), 1–12. http://doi.org/10.3389/fpsyg.2013.00863

Lee, J. F. (1986). On the Use of the Recall Task to Measure L2 Reading Comprehension. *Studies in Second Language Acquisition, 8(2)*, 201-212. doi:10.1017/s0272263100006082

Marian, V., & Fausey, C. M. (2006). Language-dependent memory in bilingual learning. *Applied Cognitive Psychology*, *20*, 1025–1047. http://doi.org/10.1002/acp.1242

Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General*, *129*(3), 361–368. http://doi.org/10.1037//0096-3445.129.3.361

Mat Daud, N. S., Mat Daud, N., & Abu Kassim, N. L. (2005). Second language writing anxiety: cause or effect? *Malaysian Journal of ELT Research*, *1*(1).

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. http://doi.org/10.1080/09658210500365763

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research, 43(6)*, 1048-1063.

Metcalfe, J. (2011). Desirable Dificulties and Studying in the Region of Proximal Learning. In A. S. Benjamin (Ed.), *Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork* (pp. 1–27). Psychology Press.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82. http://doi.org/10.1353/cml.2006.0049

Nation, P., & Beglar, D. (2007). A vocabulary size test. *JALT*, *31*(7), 9–12.

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(S1), 97–126.

Roussel, S., Joulia, D., Tricot, A., & Sweller, J. (2017). Learning subject content through a foreign language should not ignore human cognitive architecture: A cognitive load theory approach. *Learning and Instruction*, *52*, 69–79. http://doi.org/10.1016/j.learninstruc.2017.04.007

Schank, R. C. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, *3*(4), 552–631. http://doi.org/10.1016/0010-0285(72)90022-9

Schank, R. C. (1980). Language and memory. *Cognitive Science*, *4*, 243–284. http://doi.org/10.1016/S0364-0213(80)80004-8

Schoonen, R., Van Gelderen, A., De Glopper, K., Hulstijn, J., Snellings, P., Simis, A., & Stevenson, M. (2002). Linguistic knowledge, metacognitive knowledge and retrieval speed in L1, L2 and EFL writing. *Studies in Writing, Volume 11: New Directions for Research in L2 Writing*, *11*(2002), 101–122.

Sellers, V. D. (2000). Anxiety and reading comprehension in Spanish as a foreign language. *Foreign Language Annals*, *33*(5), 512–520. http://doi.org/10.1111/j.1944-9720.2000.tb01995.x

Social, T. O. &. (2012). *Europeans and their Languages*. *Special Eurobarometer 386*. Brussels.

Tulving, E., & Thomson, D. M. (1973). Encoding Specificity and Retrieval Processes in Episodic Memory. *Psychological Review*, *80*(5), 352–373.

van Weijen, D., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2009). L1 use during L2 writing: An empirical study of a complex phenomenon. *Journal of Second Language Writing*, *18*(4), 235–250. http://doi.org/10.1016/j.jslw.2009.06.003

Vander Beken, H., & Brysbaert, M. (2018). Studying texts in a second language: the importance of test type. *Bilingualism: Language and Cognition*, *21*(5), 1062–1074. http://doi.org/10.1017/S1366728917000189

Vander Beken, H., Woumans, E., & Brysbaert, M. (2018). Studying texts in a second language: No disadvantage in long-term recognition memory. *Bilingualism: Language and Cognition*, *21*(4), 826–838. http://doi.org/10.1017/S1366728917000360

Wächter, L. B., & Maiworm, F. (2014). *English-Taught Programmes in European Higher Education. The State of Play in 2014* (Vol. ACA Papers). Bonn: Lemmens.

Watcharapunyawong, S., & Usaha, S. (2013). Thai EFL students' writing errors in different text types: The interference of the first language. *English Language Teaching*, *6*(1), 67–78. http://doi.org/10.5539/elt.v6n1p67

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.

Whalen, K., & Ménard, N. (1995). L1 and L2 writers' strategic and linguistic knowledge: A model of multiple-level discourse processing. *Language Learning*.

Appendix A: Texts of experiment 2

The experiments of Zajonc and colleagues on the perception of emotions.

Scientific research on consciousness started when researchers discovered empirical evidence for the existence of unconscious processing. One of the first experiments proving that humans can be influenced by stimuli they do not perceive consciously was published by Kunst-Wilson and Zajonc (1980). The experiment consisted of two phases. In the first phase, participants were asked to watch a screen and try to discern what was presented. Ten irregular polygons (*translation: veelhoeken*) were presented five times for 1 millisecond, too short to be seen by the participants (all they saw were light flashes). In the second phase, participants were shown two polygons and had to decide which one they thought they had seen in phase 1 and which one they liked most. One of the polygon pairs had been presented in the first phase, the other was new. As expected, the participants could not indicate which polygon had been presented in the first phase (because they were not aware of having seen them). However, the participants more often than predicted by chance preferred the polygon shown in the first phase. This was the first strong evidence that emotional responses could be based on unconscious information processing.

Shortly after the study of Kunst-Wilson and Zajonc (1980), Marcel (1983) presented evidence that cognitive processing could be unconscious as well. He made use of a technique known as semantic priming. In this technique two stimuli are presented immediately after one another: the prime and the target. The usual finding is that the target is recognised faster when it succeeds a semantically related prime than when it succeeds an unrelated, neutral prime. So, the target word 'boy' is recognised faster after the prime word 'girl*'* than after the prime word 'goal'. In Marcel's experiments, target word recognition time was measured by means of a lexical decision task. In this task participants have to decide on each trial whether a presented string of letters is a word (e.g. boy) or not (e.g. doy). The target stimuli (both words and non-words) were preceded by primes to which the participants did not have to respond. In a first condition, Marcel presented the primes long enough for them to be clearly visible. In this condition, as expected he found a nice semantic priming effect. That is, participants indicated faster that boy was an existing English word if it had been preceded by the prime 'girl*'* than if it had been preceded by the prime 'goal'. In a second condition, Marcel limited the presentation time of the primes to a few milliseconds, so that participants could no longer see them consciously. Still he found a priming effect that was nearly as strong as the effect with the clearly visible prime. This indicated that the prime word did not have to be perceived consciously in order to be processed and to influence the subsequent recognition of the target word.

De proeven van Zajonc en collega's over de perceptie van emoties.

Wetenschappelijk onderzoek naar het bewustzijn is gestart toen onderzoekers evidentie vonden voor het bestaan van onbewuste informatieverwerkingsprocessen. Een van de allereerste betrouwbare proeven die aantoonden dat we door stimuli beïnvloed kunnen worden zonder ons hiervan bewust te zijn, werd gepubliceerd door Kunst-Wilson en Zajonc (1980). In de eerste fase van hun proef toonden de auteurs proefpersonen tien onregelmatige veelhoeken. Deze veelhoeken werden vijf maal aangeboden gedurende telkens 1 milliseconde (duizendste van een seconde). Na deze fase kregen de proefpersonen tien paren van veelhoeken te zien en moesten ze aangeven welke veelhoek ze in de eerste fase te zien gekregen hadden. Geen van de proefpersonen kon dit. Toen dezelfde paren van veelhoeken getoond werden en de proefpersonen gevraagd werd welke veelhoek ze het liefst zagen, bleken de proefpersonen echter vaker dan verwacht de veelhoek uit de eerste fase eruit te halen. Hoewel de proefpersonen dus geen besef (bewustzijn) hadden van de veelhoeken die in de eerste fase aangeboden waren, waren ze er in hun gedrag toch door beïnvloed.

Een andere vroege studie die het bestaan van onbewuste perceptie aantoonde,werd in 1983 door Anthony Marcel gepubliceerd. Hij werkte met een techniek die bekendstaat als *semantische priming*. Daarbij herkent de proefpersoon een doelwoord sneller als het op een semantisch gerelateerd woord (prime) volgt dan als het na een niet-gerelateerd, neutraal woord wordt aangeboden. Proefpersonen kunnen bijvoorbeeld sneller het woord 'stoel' herkennen wanneer ze voordien het woord 'tafel' verwerkt hebben. In het experiment van Marcel moesten de proefpersonen enkel aanduiden of de tweede stimulus een woord was of niet (ze moesten dus geen reactie geven op de prime). De helft van de stimuli waren bestaande woorden (bijv. stoel, hoed), de helft waren niet-woorden (bijv. stoek, loed) en de proefpersoon moest een lexicale beslissing nemen (is dit een woord of niet?) door op de linker- of de rechterknop te drukken. Eerst bood Marcel de primewoorden duidelijk zichtbaar aan. In deze conditie stelde hij zoals verwacht vast dat de proefpersonen sneller konden beslissen dat de tweede stimulus een bestaand woord vormde, wanneer het eerste en het tweede woord qua betekenis met elkaar verwant waren (tafel-stoel) dan wanneer er tussen de twee woorden geen betekenisverband bestond (boter-stoel). Vervolgens beperkte Marcel de presentatietijd van de primes zodanig dat de proefpersonen ze niet meer konden identificeren. Toch bleef het primingeffect nagenoeg even sterk.

Myth busting: Is unconscious processing dangerous?

When the first experimental evidence for unconscious information processing was published, it received quite a lot of attention in the media, because many people tended to be wary of information processing beyond their conscious control. This was partly due to Freud's claims that the unconscious is a dark force, aiming at instant gratification of its sexual and aggressive desires without regard for social or ethical considerations, which constantly tries to control humans and has to be restrained by the ego. Another reason why people did not like the idea of unconscious processing was that several urban legends about the powers of unconscious information processing were around. One of these legends was that it is possible to manipulate people's actions through subliminal advertising. Another was that unconscious messages, intermixed in music or sea sounds, can be used to heal. Still another was that hidden backward messages in songs can take control of the listeners and, for instance, incite them to commit murder or suicide.

Psychologists have been unable to find empirical support for any of the above strong claims (see Greenwald 1992; Kreiner et al. 2003; Loftus and Klinger 1992; Mayer and Merckelbach 1999). For instance, Greenwald et al. (1991) examined the effects of 'subliminal messages' (i.e. messages below the consciousness threshold) in records that otherwise sounded like normal soothing sounds. According to the makers, some records were good for improving memory; others were good for improving self-esteem. More than two hundred subjects were selected through an advertisement that contained a call for participation in a memory and self esteem improvement experiment. Greenwald et al. gave half of their participants a record to improve their memory and half a record to increase their self-esteem (this was clearly indicated on the record). Participants listened for a month at least once a day to the records. At the end of the study, they completed questionnaires about their memory performance and their self-esteem (they had done the same at the beginning of the study).

As predicted by the makers of the tapes, the participants who had listened to the self-esteem enhancing records reported higher self-esteem, and the participants who had listened to the memory enhancing records reported better memory skills. However, unknown to the participants, Greenwald et al. had changed the labels of half of the records, so that half of the participants who thought they were listening to self-esteem enhancing messages, actually heard memory enhancing messages. Similarly, half of the participants who thought they were listening to memory enhancing messages, in reality were exposed to self-esteem enhancing messages. Greenwald et al. found no difference whatsoever between the types of the actual records used; they only obtained an effect of the type of message the participants thought they had been listening to. On the basis of these findings, Greenwald et al. concluded that the positive effects participants reported were due to a placebo effect (participants expected to do better after the treatment), and not to the actual 'messages' they had been hearing. This finding agrees with the limited results of therapies based on subliminal messages.

Mythes over onbewuste processen

Onbewuste processen hebben een negatieve bijklank, omdat ze lijken te suggereren dat mensen kunnen worden beïnvloed zonder dat ze daar enige controle over hebben. Voor een deel komt dit omdat er in onze cultuur een psychoanalytisch geïnspireerd beeld heerst van een duister, seksueel beladen onbewuste, dat continu op de loer ligt om ons functioneren over te nemen en dat door het ego onder controle gehouden moet worden. Een andere reden waarom sommige mensen aan onbewuste processen magische gaven toeschrijven, is dat die processen de rationaliteit van de mens lijken te ondergraven. Een persoon die geen controle meer heeft over zijn of haar daden, is een gestoorde persoon. Tot slot doen er allerhande mythes de ronde over ziekteverwekkende en helende invloeden van stimuli die niet bewust waargenomen kunnen worden. Voordat we het kunnen hebben over de interacties tussen bewuste en onbewuste processen, moeten we dus eerst kijken naar wat er van deze overtuigingen waar is.

Een klassieke studie over subliminale invloeden werd uitgevoerd door Greenwald et al. (1991). Zij onderzochten de doeltreffendheid van zelfhulpcassettes, die in die tijd een rage waren. Op deze banden waren kalmerende geluiden te horen samen met ‘subliminale boodschappen’ die volgens de producenten de luisteraar ertoe aanzetten om gewicht te verliezen, een beter geheugen te krijgen, te stoppen met roken of een gunstiger zelfbeeld te krijgen. Meer dan tweehonderd proefpersonen werden geworven door middel van een advertentie, waarin opgeroepen werd tot deelname aan een experiment ter verbetering van het geheugen of de zelfachting. Eerst werd aan de proefpersonen gevraagd om een vragenlijst over hun zelfbeeld en hun geheugenprestaties in te vullen. Daarna kregen ze een bandje mee dat volgens de fabrikant ofwel aanzette tot een beter geheugen ofwel tot een hogere zelfachting (dit was op het bandje duidelijk aangegeven). Aan de proefpersonen werd gevraagd om hier gedurende een maand elke dag minstens 1 keer naar te luisteren. Aan het einde van de periode vulden de proefpersonen opnieuw een vragenlijst in over hun zelfbeeld en hun geheugenprestaties.

Uit de resultaten bleek dat de mensen die een bandje gekregen hadden ter verbetering van hun geheugen effectief een hogere geheugenefficiëntie rapporteerden en dat mensen die een bandje gekregen hadden ter verhoging van hun zelfachting, hier ook een duidelijk effect van ondervonden. Wat Greenwald et al. echter niet aan hun proefpersonen verteld hadden, was dat slechts de helft van de proefpersonen het bandje gekregen had dat op het etiket stond. Bij de andere helft van de proefpersonen hadden de onderzoekers de etiketten verwisseld. Uit de resultaten bleek dat er geen verschil bestond tussen het‘effect’ gemeld door de proefpersonen die naar het juiste bandje geluisterd hadden en het ‘effect’ gemeld door de proefpersonen die naar het verkeerde bandje geluisterd hadden. Het geheugen was evenveel verbeterd bij de proefpersonen die naar een bandje ter bevordering van de zelfachting geluisterd hadden als bij de proefpersonen die naar een bandje ter bevordering van het geheugen geluisterd hadden. Op basis van deze bevindingen besloten Greenwald et al. dat zelfhulpbandjes met subliminale boodschappen mensen inderdaad een beter gevoel geven, maar dat dit niet te danken is aan de subliminale boodschappen, maar aan een placebo-effect, veroorzaakt doordat de personen verwachten dat ze beter zullen worden door het luisteren naar de bandjes.

Appendix B: Tests of experiment 2 (English version)

Zajonc correction key

1. What does Zajonc and colleages' experiment prove? (/2)

   - It proves that our behaviour can be affected
   - By things we have not consciously perceived

2. What is semantic priming? Explain (/2)

   - The finding that a target word is recognised faster
   - When it follows a semantically related prime than when it follows an unrelated word (prime – target /0.5; semantic relation /0.5)

3. Which manipulation in Marcel's (1983) experiment was crucial to the conclusion about unconscious processing? (/1)
   - The manipulation of presentation time: limiting the time in the second experiment so the participants could no longer identify the primes (masked priming)

Myths correction key

1. Based on the text you have just read, do you think that self-help books against stress work, and why/why not? (/2)

   - Yes it probably works
   - Since people who chose to read these books, are looking for a solution to a problem and probably believe in this solution. Similar to the study mentioned in the text, their expectations would lead to a placebo effect. (Note: if a participant answers no but explains that, if anything happens, it is because of the beliefs of the reader, this counts as a correct explanation)

2. Do you think the same results would be found in a group of participants who must participate in an experiment as part of study requirements, and why? (/2)

   - No
   - The placebo effect cannot be as strong: they do not participate voluntarily because they believe it helps

3. Which field of psychology was part of the reason people feared unconscious processes? (/1)

   - Psychoanalysis / Freud