Text comprehension and memory in L1 and L2

Heleen Vander Beken

Supervisor: Prof. Dr. Marc Brysbaert
Co-supervisor: Prof. Dr. Robert Hartsuiker

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of Psychology

Academic year 2017–2018
Text comprehension and memory in L1 and L2

Heleen Vander Beken

Supervisor: Prof. Dr. Marc Brysbaert
Co-supervisor: Prof. Dr. Robert Hartsuiker

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Doctor of Psychology

Academic year 2017–2018
Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt.

– Ludwig Wittgenstein

What’s in a name? A rose by any other name would smell as sweet.

– Juliette in Romeo and Juliette, William Shakespeare
# Table of Contents

ACKNOWLEDGMENTS .......................................................................................................................... 7

CHAPTER 1: INTRODUCTION .................................................................................................................. 11

1. **An Educational Perspective: English as a Medium of Instruction** ........................................... 14
   1.1 English as Medium of Instruction Worldwide ............................................................................ 14
   1.2 State of affairs in the Netherlands and Flanders ...................................................................... 16
   1.3 Insufficient focus on cognitive consequences .......................................................................... 17

2. **A Theoretical Perspective: Bilingual Memory for Texts** ............................................................. 17
   2.1 Semantic memory ......................................................................................................................... 17
   2.2 Bilingual memory ....................................................................................................................... 18
      2.2.1 Models for bilinguals visual word recognition ...................................................................... 19
      2.2.2 Language non-selective frameworks explaining L2 disadvantages ...................................... 22
      2.2.3 Predicted and reported disadvantages in L2 ..................................................................... 23
      2.2.4 A conceptual base for thought ............................................................................................... 24
      2.2.5 The encoding-specificity principle ......................................................................................... 25
   2.3 The relation between reading comprehension and memory ....................................................... 29
      2.3.1 The reading comprehension process .................................................................................... 29
      2.3.2 Factors Predicting Reading Comprehension .......................................................................... 30
      2.3.3 Reading comprehension in L2 .............................................................................................. 32
      2.3.4 Pitfalls in reading comprehension assessment ..................................................................... 33
      2.3.5 Text accessibility ................................................................................................................... 34

3. **The Current Dissertation** .............................................................................................................. 35
   3.1 Some Methodological considerations .......................................................................................... 35
      3.1.1 Comparing languages ............................................................................................................ 35
      3.1.2 Power ..................................................................................................................................... 36
      3.1.3 Excluding other factors ......................................................................................................... 36
   3.1 Research lines and chapters ........................................................................................................ 37

REFERENCES ........................................................................................................................................... 40
# Table of Contents

CHAPTER 2: THE IMPORTANCE OF TEST TYPE ................................................................. 49
  ABSTRACT .................................................................................................................. 51
  INTRODUCTION ....................................................................................................... 52
  METHOD .................................................................................................................. 56
  RESULTS ................................................................................................................ 63
  DISCUSSION .......................................................................................................... 70
  REFERENCES ......................................................................................................... 76

CHAPTER 3: NO L2-DISADVANTAGE IN LONG TERM RECOGNITION MEMORY ......... 81
  ABSTRACT ................................................................................................................ 83
  INTRODUCTION ..................................................................................................... 84
  METHOD ................................................................................................................ 88
  RESULTS .............................................................................................................. 93
  DISCUSSION ....................................................................................................... 100
  REFERENCES ..................................................................................................... 104

CHAPTER 4: A FURTHER INVESTIGATION OF FACTORS INVOLVED IN THE L2 COSTS ........... 107
  ABSTRACT ............................................................................................................. 109
  INTRODUCTION ................................................................................................... 110
  EXPERIMENT 1: CUED RECALL OF HANDBOOK PARAGRAPHS IN L1-L1, L2-L1, L2-L2 CONDITIONS ........ 114
    Method ............................................................................................................. 114
    Results ........................................................................................................... 120
    Discussion .................................................................................................... 122
  EXPERIMENT 2: CUED RECALL AND RECOGNITION OF AN ACADEMIC ARTICLE .................. 123
    Method ............................................................................................................. 123
    Results ........................................................................................................... 126
    Discussion .................................................................................................... 129
  GENERAL DISCUSSION ..................................................................................... 130
  REFERENCES ..................................................................................................... 138

CHAPTER 5: THE L2 RECALL COST IS NOT JUST A PRODUCTION EFFECT ......................... 141
  ABSTRACT ............................................................................................................. 143
# Table of Contents

**REFERENCES** ........................................................................................................ 220  

**ENGLISH SUMMARY** .......................................................................................... 223 

**NEDERLANDSTALIGE SAMENVATTING** .............................................................. 230 

**APPENDICES** ......................................................................................................... 239 

**APPENDIX A: BIOLOGY TEXT/TESTS (CHAPTER 2, 3 AND 5)** .............................. 241  
  Appendix A.1 Biology texts used in Chapter 2, 3 and 5 ........................................ 241  
    *Appendix A.1.1: Texts about The Sun and the Sea Otters* .................................. 241  
    *Appendix A.1.2: Vocabulary support given in Chapter 5* .................................. 245  
  Appendix A.2 Tests for the biology texts ............................................................. 248  
    *Appendix A.2.1: Tests used in Chapter 2* ...................................................... 248  
    *Appendix A.2.2: Tests used in Chapter 3 (parallel versions)* ......................... 255  
    *Appendix A.2.3: Tests used in Chapter 5* ...................................................... 260  

**APPENDIX B: PSYCHOLOGY TEXTS/TESTS (CHAPTER 4)** .................................... 263  
  Appendix B.1 Chapter 4 Experiment 1 texts ...................................................... 263  
  Appendix B.2 Chapter 4 Experiment 1 tests ...................................................... 267  
  Appendix B.3 Chapter 4 Experiment 2 text ....................................................... 268  
  Appendix B.4 Chapter 4 Experiment 2 test ....................................................... 276  

**APPENDIX C: PROFICIENCY TESTS** ..................................................................... 284  
  Appendix C.1 Dutch multiple-choice vocabulary test ......................................... 284  
  Appendix C.2 Spelling tests .................................................................................. 286  

**APPENDIX D: QUESTIONNAIRES** ..................................................................... 288  
  Appendix D.1 Motivation and text questionnaire Chapter 2 .............................. 288  
  Appendix D.2 Motivation questionnaire Chapter 3 ............................................ 289  
  Appendix D.3 Text questionnaire in Chapter 3 and 5 ........................................ 293  
  Appendix D.4 Additional set of questions in motivation questionnaire of Chapter 5  
    Experiment 2 ........................................................................................................ 294  

**DATA STORAGE FACT SHEETS** .......................................................................... 295
Acknowledgments

Even in science, our only certainty is that we have no certainty.

But it is a certainty that this dissertation could not have come about without the help of many kind people on a wobbly four-year road to this finish line.

Though they were not directly involved in the work presented here, I would first like to express my gratitude to the following two inspiring academics.

During my literature and linguistics studies, I enjoyed the enthusiastic and engaging courses taught by Professor Dominiek Sandra (UA). Dominiek, you awakened my curiosity and planted the first little seed of scientific inquiry in my mind; you challenged me to think critically and took the time for elaborate brainstorming about my first research design ever. It is still a pleasure to run into you on academic events, because our conversations leave me with a sense of encouragement and mutual respect.

In addition, I would like to write a few words in memoriam of Professor Peter Mariën. I had the opportunity to follow his work in the ZNA Middelheim hospital for a couple of days. I was impressed by Peter’s elaborate knowledge, his seemingly effortless and swift analytic ability, his respectful and empathic attitude towards patients, and his kindness towards people in general. During that internship I learnt a great deal about neurolinguistics, guided by – as we would say in Dutch – een crème van een mens.

After these first encounters with great academics, I had the privilege to work with researchers I cited in my masters (I assume this is a correlation, not a causal relation), such as my supervisor Marc Brysbaert. I vividly remember the moment right after my job interview: when I left the room, I suddenly realised that I had been talking to anyone but the person who would be hiring me... He had just been silently observing how I tried to sell myself. Marc, thank you for giving me this chance and for counting my successes rather than my failures. Your door was always open, right across the corridor, and I must have knocked on it a zillion times. You took up to two and half hours (!), even on a Friday afternoon, to debate an approach with me that was mainly based on my limited knowledge of statistics at the time. And I must say: if there were a Guinness book of
academic records, your name would be under the speed record for reviewing manuscripts.

I would also like to thank my co-supervisor Robert Hartsuiker for his detailed feedback on this dissertation, and Professors Wouter Duyck and Martin Valcke for the valuable input during project and advisory committee meetings. Merci à vous aussi, Fabienne Chetail, d’avoir pris le temps de me conseiller chaque année. Evy Woumans, thank you for following up on my work, but even more so for contributing to it and for sharing a room and vegan restaurants with me in Granada. In addition, I am indebted to Evelyne Lagrou and Michaël Stevens for their help in the data collection and analysis of the first empirical chapter.

There is a long list of colleagues who contributed to my work or to a pleasant working environment, too many to name. But I am especially thankful the following colleagues of the LEMMA-project I was part of.

Ellen, thanks for being my partner in crime in Vienna and for welcoming me on my many visits to the second floor (together with your lovely office mates). I remember the good laughs we had during our marathon writing sessions and meetings, which we sometimes left exhausted, but often full of ideas. I admire your excellent communication to external partners and your willingness to listen to my suggestions.

Nicolas, it was definitely a pleasure to take part in one of your research papers. Apart from that, I really enjoyed your company at conference dinners, at work, and abroad. Thanks for the thoughtful and encouraging advice throughout our collaborations and even in the last days before my deadline. I hope that we can conquer our arachnophobia together at some point in the future and that I will taste your delicious homemade guacamole again.

Wouter, I remember how we met when we both were a little lost in the administration department. Thank you for making the first moments at a new job (and our other ventures) less awkward with your reassuring presence. I know few people who are so dedicated to their job and their family. It’s just a pity that you wouldn’t let us hear more of your lovely accent...

Aster, congratulations on courageously tackling the combination of a PhD in Ghent and a family in Rotterdam. It was always nice to have you around, and I very much
appreciated your ever supportive attitude and bright mood. I hope we can catch up some
time, so I can hear all about the awesome job you will hopefully find.

Last but not least, Toru, I couldn’t have done the word list recall study without your help. Marcelo, I am glad you were on our team as well. I wish you both the best of luck in your future adventures in Brazil, Japan or elsewhere.

Sometimes there need not be a professional reason to become a team. I felt very much at home in our little enclave on the fifth floor (de vier van ’t vijfde). Helena, I have always appreciated your enthusiasm, your passion for neuroscience, and, now and then, your up-front and justified criticism. Maaike, thanks for the many hilarious moments, candid conversations, and for your support and advice. Lise, I am glad we have gotten to know each other better and better during our last-minute scientific and cultural excursions. Thank you for sharing tips and tricks and your passion for science communication with me.

Research also comes with difficult technicalities. I am thankful to Toon De Pauw for developing a beautiful website (and Pawel for providing a server), and of course, to Lies and Christophe, who keep the department running every single day.

My fondness of language is definitely inherited or, at the least, the consequence of my upbringing. Next to silly puns, my gene pool mostly offered opportunities. I will always be indebted to my parents for those opportunities and for believing in me. In addition, I am blessed with wonderful siblings who have my back and a cheerful little niece reminding me of what is important. Furthermore, thanks to Tine for her wisdom.

Finally, having fun with friends is indispensable. Jana, I feel huge respect for your intelligence and diligence and find it a privilege to work or to play music with you. I am grateful to Céline as well, for a lifelong friendship and for her ability to put things into perspective. My appreciation also goes to Annemie for putting up with my frustrations, to Julie, Karen, Eva, and Astrid, for providing a safe haven, to Ilke for patiently caring and to many more for lifting my spirit.
INTRODUCTION
People gather knowledge mostly through language, at least from a certain age. Hence, language is a necessary means for us to describe our world and at the same time influences the way we see that world. An example is the way we observe colour: depending on the language we speak, we categorise colours differently (Davidoff, Davies, & Roberson, 1999). The fundamental question here is what comes first: the language or the concepts behind it. Does language only illustrate our thoughts, or does language itself shape the way we think? These questions have been subject to speculation and research for decades, with few to no conclusions carved in stone (some recent examples are E. Davis, 2015; Lupyan, 2016).

The situation becomes more complex when a second language is involved: what aspect of which language influences our thinking? Can any evidence be found in bilingual subjects for or against a conceptual, language-independent base for thought? In addition, new questions arise. Do skills in one language suffer from the other language? Is the flexibility of switching between languages a cognitive asset or a burden to the brain’s functionality? To what extent are the languages separate entities in our mind and how do we cope with that?

The last question comes the closest to the topic of the current dissertation. In a series of experiments, it is investigated how information offered through texts is encoded, stored and retrieved, and most importantly, whether these processes are impaired because of the use of a non-native language. This question is of utmost practical importance in a globalising world, but the results can also reflect the processes behind memory and bilingualism, by adding new elements to a large body of experimental evidence. In this dissertation, bilingual memory is investigated using quantitative psycholinguistic methods. In the present introduction, we will describe the current state of affairs concerning English as a Medium of Instruction, and explore the theoretical frameworks from different disciplines that are related to our research question. As pointed out by Francis and Gutiérrez, no “established theory of bilingual proficiency and memory” is available, we use ideas from other domains of bilingual research and from other domains of memory research to make predictions about bilingual memory.
1. An educational perspective: 

English as a Medium of Instruction

1.1 English as a medium of instruction worldwide

It is estimated that about half of the world population is bilingual (Grosjean, 1989; TNS Opinion & Social, 2012). In this context, we will define bilinguals not as people who were raised with two native languages or who master two languages to an equal extent (this is referred to as balanced bilinguals), but as anyone who masters more than one language to a certain extent (Grosjean, 1989). Especially in this sense, English is clearly a world language. Reports requested by the European Commission indicate that more and more Europeans know more than one language (TNS Opinion & Social, 2012), and that English has become widespread both in a geographical sense and across domains. For example, technology is a domain in which English is very dominant, but this also holds for science. With this general importance of English and its status as a lingua franca for research, English is slipping into higher education as a medium of instruction (EMI). Globalisation in general, resulting in increasing mobility for students, amplifies the presumed need for English-language education. EMI can be defined as ‘The use of the English language to teach academic subjects in countries or jurisdictions where the first language (L1) of the majority of the population is not English’ (Dearden, 2014, p.2). Though it is sometimes referred to as CLIL (Content and Language Integrated Learning) or as part of CLIL, it is different in the sense that language learning is not a goal in itself in most of the situations that are referred to in this dissertation.

This phenomenon has come up very quickly in Europe over the last two decades. Wächter and Maiworm reported in 2014 that (fully) English taught programmes at European higher education institutions increased by 1000% in the years between 2001 and 2014 (though some of this growth might have been due to underrepresentation in previous surveys). In contrast with this extreme growth, the number of students this relates to is rather limited. Following the authors’ estimation, this would only reflect
about 1.3% of total student enrolment in the countries that were included in the survey, and overall about one fourth of all institutions offer an English taught programme (mind that this does not include programmes partially taught in English). But this evolution has spread wider than Europe, with EMI emerging at universities in for instance Israel (Inbar-Lourie & Donitsa-Schmidt, 2013), China (Hu, Li, & Lei, 2014), Taiwan (Han & Singh, 2014), Japan (Galloway, Kriukow, & Numajiri, 2017; Rose & McKinley, 2018).

1.2 State of affairs in The Netherlands and Flanders

Based on three indicators (the number of higher education institutions providing ETPs, the share of ETPs in comparison to all programmes, and the enrolment in ETPs compared to total national enrolment), the Netherlands is ranked as the ‘leader’ in providing English-taught programmes among 28 European non-English speaking countries (Wächter & Maiworm, 2014). In that same division, Belgium is only ranked 17th (though no distinction has been made between Flanders and Wallonia here, so the numbers are possibly higher for Flanders). Twenty percent of bachelor degrees in the Netherlands are offered in English, with another 10 percent offering the choice between Dutch and English. For master degrees, this is 59% (note that these numbers vary greatly depending on the subject), and for colleges (“hogescholen”) it is 8% (KNAW, 2017). In Flanders, this was only 22% for masters and under 2% for bachelor degrees in 2015 (VLOR, 2017) As for the number of students, one fifth of students at universities in the Netherlands are international students, mostly from countries in which English is not an official language. Of all registered students in the Netherlands in general, 41% are registered in an ETP according to the 2017 KNAW report. To sum up: the matter of foreign-language education is clearly of a very different size in both countries. Nonetheless, it is a recurrent topic of debate at universities and in newspapers in both countries.

Another way to compare these areas is by law. Dutch law prescribes that education must be given in Dutch, except in the following three exceptions: (1) when the subject of the programme is another language, (2) when a course is taught by a non-Dutch speaker,
(3) when the specific nature, the set-up, or the quality of education, or the origin of the students requires the use of another language, in accordance to the code of conduct of the institution’s governing body (as reported by KNAW, 2017; our translation). In reality, the last rule is used in abundance, as the numbers above have shown, with limited motivation for the exception (KNAW, 2017). In addition, there is a statutory duty which obliges universities to take up the task of improving Dutch oral and writing proficiency (the skill to “express”, literally). In Flanders, law is much more concrete, specifying certain conditions and numbers. For ETPs, the maximum allowed degrees are at 6% and 35% (VLOR, 2017), and they are only allowed when the added value is proven and when an alternative is available in Dutch (amongst other rules). Interestingly, this rule is debated intensively, while the actual numbers are far below this decretal maximum.

Of course, programmes can consist of a mixture of English and Dutch courses as well. In Flanders, 18.33% of the courses within a degree can be in English, for masters this is 50%. In addition, English handbooks or academic literature can also be used in Dutch courses, which is not limited by law nor counted in these numbers. For example, about half of the list of handbooks that psychology students need at Ghent University is in English (Dirix et al., 2017). In other words: more English is used than these numbers or rules indicate, especially for written materials.

1.3 Insufficient focus on cognitive consequences

The discussion on EMI usually pivots around ideological issues such as the status of Dutch as an academic language, inclusiveness of education, and the economic reasons of choosing for English (see for example KNAW, 2017, for an overview of arguments pro and con; Wilkinson, 2013). Research in several geographical areas points out possible issues such as increasing inequality (Hu et al., 2014; Inbar-Lourie & Donitsa-Schmidt, 2013), disruption of local language ecologies and the threat to the principal language of instruction (e.. Chinese, Li, 2013), motivational issues toward the use of EMI (Han & Singh, 2014), loss of information in curriculum design (Wilkinson, 2013), etc. Though all of these can be legitimate arguments, little attention is paid to the cognitive consequences of the choice. When cognition of students is considered, it is mostly in
opinion pieces that are strongly against EMI, stating for example that “learning outcome is only one fourth” (e.g. column by Martin Sommer in de Volkskrant, 19-08-2016, our translation). Nevertheless, little is really known about how the use of a foreign language influences the knowledge of students, while this question is central to the quality of education. And if any attention is paid to it, it is mostly to English-taught classes, not to the difficulty of English study materials. For example, it has been investigated whether EMI increases proficiency in English (e.g. positive effects reported by Tatzl, 2011) and whether proficiency level of teachers and students influences understanding of the materials (see Hu et al., 2014). In this dissertation, we attempt to answer the question how well information is encoded and retained – and in which way – from English texts by students whose dominant language is Dutch. Apart from the applied importance we have just discussed, this also increases our understanding of bilingual memory for learnt information in general.

2. A theoretical perspective: bilingual memory for texts

2.1 Semantic memory

Let us first define the part or type of memory that is subject to our research. Since our interest is in cognitive effects of studying in a non-native language, the focus in this dissertation lies on the information learnt from a text. The goal of education, apart from training skills, is to let students absorb knowledge on the long term, so it can be retrieved and used in future professions. As such, an aspect of long-term memory is measured here (as opposed to short-term or working memory). Within long-term memory, a distinction is made between declarative or explicit knowledge, referring to conscious knowledge that can be verbalized (e.g. Paris is the capital of France) and procedural or implicit knowledge, which refers to the knowledge on how to perform an action, such as how to
drive a car (Brysbaert, 2014). Our interest lies in the first. Declarative knowledge is
divided further into episodic and semantic memory. Though both could possibly be
influenced by language, information learnt from texts ought to be stored in semantic
memory, which contains all our factual or world knowledge and is not tied to a specific
event, as is episodic memory (Brysbaert, Ameel, & Storms, 2014; Graves & Altarriba,
2014; Hardt, Nader, & Nadel, 2013). Nevertheless, this division can and will be
questioned on a theoretical level in the first empirical chapter in this dissertation. The
division between episodic and semantic memory corresponds to a neurobiological
difference with hippocampal memory traces containing episodic information and
neocortical traces containing semantic memory. It can be argued that all semantic
knowledge is at first episodic, and that it is transferred to semantic memory when the
hippocampal traces, and thus the episodic information, fade out (Hardt et al., 2013). Still,
other authors claim that some memories can shift between those types in both directions,
and that it is debated which comes first (Schroeder & Marian, 2014). Nevertheless, when
information is retrieved from studied texts (e.g. during an exam), this is mostly from
semantic long-term memory, which will be the focus of the following paragraphs.

2.2 Bilingual memory

A large body of research has focused on semantic memory in bilinguals, at least on
the word level. The initial assumption that a bilingual is basically the sum of two
monolinguals, with a mental lexicon for each language (the so-called monolingual or
fractional view), has been rejected years ago (Grosjean, 1989). Over the last decades,
evidence refuting this idea has accumulated (e.g. de Groot, Delmaar, & Lupker, 2000;
Duyck, Assche, Drieghe, & Hartsuiker, 2007; Kennette & Van Havermaet, 2012; Moon &
Jiang, 2011). Though different frameworks explain language processes and interlingual
interference in different ways, there is a consensus that lexical access is language-non
selective. In the following sections, some historically important models and frameworks
of bilingual memory will be discussed.
2.2.1 Models for bilingual visual word recognition

A historically influential model sets out from the assumption that there are indeed separate lexicons for each language, with a connection to a separate, abstract level for word meanings. The Revised Hierarchical Model (Kroll & Stewart, 1994, see Figure 1) thus explained language differences by the strength (or lack thereof) of connections between the lexicons and the semantic level. Though some experimental findings can indeed be sufficiently explained by this model, this account was falsified by evidence of interlingual interference, even in unilingual contexts (e.g. Van Assche, Duyck, & Hartsuiker, 2012).

![Figure 1. The Revised Hierarchical Model (Kroll & Stewart, 1994)](image)

As a reaction to this falsification, several theoretical and computational models were developed that did not include entirely separated lexicons but an integrated lexicon and language nodes (e.g. the Strong Phonological View by Frost, 1998). In addition, the importance of the context or task was emphasized by introducing a task schema or control level (e.g. the Bilingual Interactive Activation model of Van Heuven, Dijkstra & Grainger, 1998, see Figure 2). Putting these elements together, the updated BIA+ model (Dijkstra & van Heuven, 2002, see Figure 3) is currently one of the most influential...
models. In this system, language is coded as a node, separate from a semantic level, and sublexical and lexical levels are split up in orthographic and phonological elements.

**Figure 2.** The Bilingual Interactive Activation Model (BIA, van Heuven, Dijkstra, & Grainger, 1998)
Figure 3. The BIA+ model (Dijkstra & van Heuven, 2002)
In short, it is agreed upon that both languages of a bilingual are active during processing, even when only one language is needed. This does not only hold for word recognition, but for sentence reading as well (for a survey, see Van Assche, Duyck, & Hartsuiker, 2012). Furthermore, in current models of bilingual language processing, semantic concepts are supposed to be shared between languages (for a review of the models that were discussed, see Brysbaert & Duyck, 2010).

2.2.2 Language non-selective frameworks explaining L2 disadvantages

To our knowledge, no theoretical or computational models (like the BIA+) have been developed for higher level processes in bilinguals, though reading or studying a text is a more ecologically valid task than word recognition in isolation for example. Nevertheless, there are broader frameworks within the language-non-selective access view which predict L2 disadvantages. A first account sets out from interference between the first and the second language. This cross-linguistic interference (Weber & Cutler, 2004) hypothesis assumes competition between representations of both languages at the lexical level which interferes with recognition. Related to memory processes, this framework could also explain what is called retrieval-induced forgetting (Levy, Mcveigh, Marful, & Anderson, 2007; Runnqvist & Costa, 2012): the finding that retrieving a concept in one language hinders retrieving it in the other language (other authors disagree; Runnqvist & Costa, 2012). When reading a text in a second language, this interference would hinder and slow down the encoding process, which is what you would intuitively expect. Another possible reason for slower and hampered L2-reading is given in the weaker-links hypothesis (also called the frequency-lag hypothesis; Gollan et al., 2011). This framework predicts the same issues for L2 words as for low-frequent L1-words. Those words are less familiar and their semantic representations are less detailed (Finkbeiner, 2002). The lower familiarity could lead to better recognition memory for words, since less familiar words are more unique in memory (W. S. Francis & Gutiérrez, 2012). The third and last hypothesis assigns L2 disadvantages to a different prerequisite of reading: working memory capacity. The resource hypothesis presumes that working memory is taken up more in a second language, leaving less capacity for higher order processing such as integrating prior and new information and monitoring comprehension (Sandoval,
Gollan, Ferreira, & Salmon, 2010). If the cognitive load is indeed higher, a compensation strategy could be used in which readers invest more effort in the task, resulting in a seemingly contradictory better memory performance. This is one of the reasons all experimental tasks in this dissertation include a time limit.

2.2.3 Predicted and reported disadvantages in L2

The frameworks discussed in the previous section are not always mutually exclusive: some findings can be explained by a combination of frameworks (e.g. resource and weaker-links hypothesis, (Sandoval et al., 2010). And while they would yield different effects in lower-order processes, it seems they would lead to similar L1-L2 effects on higher-order processes. One such effect is that reading rate would be slower in L2. And indeed, L2 reading spans (the number of sentences you can process whilst remembering the last word of every sentence) are reported to be shorter than L1 reading spans, which indicates that L2 processing is indeed more demanding than L1 processing, (Vejnović, Milin, & Zdravković, 2010; as reported by Szmalec, Brysbaert, & Duyck, 2014). In a study using some of the materials presented in this dissertation, it was found that reading or studying a text in a second language took about 18% longer than in a native language (Dirix, Vander Beken, De Bruyne, Brysbaert, & Duyck, under review). Less capacity or weaker links could also lead to diminished understanding (since less inferences are made or less links to semantic features are available) or more forgetting (since slower reading leads to less repetition of the text and weaker links can fade out sooner).

Empirical evidence has shown that learning tasks sometimes result in worse performance in L2 (here we are considering unilingual conditions; the effect of congruency between text and test language will be discussed in section 2.2.5). For example, Reithofer compared the cognitive end-result when an audience listened to a 15-minute speech in English as an L2 (of an Italian speaker) versus a simultaneous interpretation in L1 German (Reithofer, 2013). Better results were found for the L1 interpretation version, but several points of criticism cast doubt on this result. For example, the foreign speaker had a heavy accent different from that of the listener (bilingualism research has shown that a foreign accent decreases perceived reliability, Lev-Ari & Keysar, 2010), no attention was paid to speaker characteristics (different
speakers in different conditions), the interpreter might have reduced complexity, and the interpretation was edited.

If indeed there is a shared and abstract semantic level and language non-selective access, you would assume text information is retained in this form and thus no longer connected to the language input. So the quality of that what is retained will then be determined by the fluency of the processes during encoding, as in the aforementioned frameworks.

2.2.4 A conceptual base for thought

The idea of a shared and abstract semantic level, present in the language non-selective frameworks, is very similar to older theories in cognitive psychology. In 1972, Schank developed his conceptual dependency theory of natural language understanding, setting out from the premise that the base of natural language is conceptual and interlingual, and that its elements are not words but concepts (conceptual dependency theory; 1972). This idea was derived from the observations that people did not remember the surface structure of a text but the main ideas in it and that bilinguals could easily switch between languages. In other words:

... because people could easily translate from one language to another and, in a sense, think in neither, there must be available to the mind an interlingual, i.e., language-free, representation of meaning. (Schank 1980, p.244)

In the same vein, Alba and Hasher explain their take on schema theory (1983). In schema theory, it is assumed that information is reduced by a process of abstraction for encoding in memory, in which the format of a message is lost. As such,

memory is abstractive in that a verbatim record is not left behind; rather, meaning appears to have highest priority for storage (1983, p. 212)

Still, as the authors point out, there were issues with this interpretation: experimental evidence (even at that time!) shows examples of verbatim memory in highly overlearned information (like prayers for religious people), but also for information that was only heard once (such as recognition for sentences used in a lecture or overheard in a
TV show). Their conclusion that “memory appears to contain far more syntactic and lexical detail” (Alba & Hasher, 1983, p. 215) contradicts the previously quoted assumption of abstraction. In our opinion, these findings could reflect episodic memory instead of semantic memory, in which case the theory is not really contradicted by these findings (see the next section for theories that argue for language-specificity mostly in episodic memory).

Though these ideas are grounded in a different domain, they are actually very close to the generally accepted bilingual mental lexicon with a language-independent semantic level. In addition, the idea that meaning is not stored in a linguistic code, is very intuitive. As put by Oakhill, Cain and Elbro, “what readers remember of a text is not the wording [...] but the meaning” (2014, p. 11).

2.2.5 The encoding-specificity principle

Still, memory research also reports a very different finding. Memory (retrieval) benefits from contextual similarities between the moment of encoding and that of retrieval. In other words: when the context of an event recurs during the recall, it will facilitate recall of that same event. This is called the encoding-specificity principle (Tulving & Thomson, 1973). A large body of evidence supports this principle. For example, when people read an article in silent or noisy conditions, they recall and recognise more in context-congruent conditions (Grant et al., 1998). The question that follows from these findings is twofold: is language a form of context that yields encoding-specificity and if so, does language-specificity affect semantic memory – when the context of encoding is no longer part of the actual memory trace? In that case, learning in L2 might be beneficial. For the first question, the answer seems to be positive, as the following evidence from three modalities demonstrates.

2.2.5.1 Autobiographical Memory is Language Dependent

Evidence for language-dependent memory has been found in autobiographical memory research. It appears that people can access memories of events more easily when recalling them in the language the event took place in. For example, when one person relates the same experience in L1 on one time and in L2 on another time, the most detailed description will be that which is told in the language of the experience itself.
(Javier, Barroso, & Muñoz, 1993; as reported by Schrauf & Rubin, 1998). However, this is limited evidence since only 5 participants were tested. Other studies have used word cues in a specific language to which (late) bilinguals or immigrants had to respond with memories. It was found that (a) the average age of memories is lower when the cues are presented in the mother tongue, (b) more memories are retrieved for cues in the mother tongue (Matsumoto & Stanny, 2006), (c) more memories are retrieved when the cue language matches the language of the recalled event (though this is not the case in highly proficient bilinguals) or of first thought (Matsumoto & Stanny, 2006; Schrauf & Rubin, 1998). Though some small differences in the results of those studies can be found, evidence in favour of language-dependent autobiographical memory accumulates. Still, this does not mean that memory is language-dependent per se. Firstly, Marian and Neisser (2000), who found similar effects, argue that the context of retrieval is confused with the cue language: context modulates the language of retrieval independently of the cue language. Secondly, autobiographical memory is episodic memory for real-life events, in which the context assumedly plays a larger role than in studying a text. As Marian and Fausey (2006) state, “it is possible that language-dependent memory manifests itself differently for different types of information” (p.1041). Therefore, we will look into memory processes that are closer to text recall.

2.2.5.2 Word List Recall

In studies of bilingual memory, researchers have offered word lists in L1 and L2 for participants to memorise. Some findings suggest that recall is not concept-mediated, but rather literal (Watkins & Peynircioglu, 1983). Watkins and Peynircioglu have let participants complete words or translation of words they had seen in a mixed language word list. They found that participants have more difficulty completing translations of words compared to their original presentations, though they still perform better on translations than on words that were not presented in either language (1983). The latter finding proves that concepts are accessed during word list recall (otherwise, translations would not have scored higher than original words). When a unilingual list is offered, though, translation completion is not better than the completion of unpresented words, which is interpreted as evidence for language-specificity (note that participants were informed that both languages could be used just before the cued recall test). Nevertheless, this task is very different from natural language processing. Furthermore,
if a unilingual list was offered, the lack of salience of the other language might inhibit the production of a translation to that language in the completion task, in which case this result is no evidence for language-specific memory.

Additional evidence for conceptual access during word list recall was reported by Glanzer and Duarte (1971). They presented mixed word lists in which concepts were repeated in the same or in a different language (English and Spanish). Participants were presented with words at a 2-second rate. They had to read the words aloud and try to remember as many words as possible after the full list. Between-language repetition yielded higher recall scores for words, suggesting that the participants remembered the concepts rather than the mere lexical representation. Once again, this experimental condition differs from natural reading, in this case because languages were intermixed.

Furthermore, Nott and Lambert (1968) found that bilingual participants recalled equal numbers of words in L1 and L2 (and mixed) word lists, except when the words from the list could be organized into specific semantic categories. It was only in categorized lists that bilinguals had a disadvantage in their weaker language. The authors interpreted the results in the following way: when no categories are present, the lexical representations of the words are studied and retrieved. However, when semantic categories are present, participants need to access the conceptual representations to benefit from the categorisation words, and this takes longer in L2. If this interpretation is correct, we might observe similar differences in text recall: to understand a text, the reader definitely needs to access concepts. Durgunoğlu and Roediger (1987) tested the effect of task demands on language-specificity and confirmed that word-fragment completion is language-specific, but free recall, a conceptually driven task, is not (which confirms the suggestion that the productive element of word fragment completion is responsible for part of the effect). In conclusion, word list recall research cannot give a straightforward prediction of how bilingual memory processes information in L2 texts.

The results above seem to suggest there language-specificity effects in recall depend on the task or strategy that is used, while recognition memory is more language-specific. In a more recent study, in which the authors investigated whether false memories (recollection of words that were not presented but are related to previously presented words) cross language boundaries, accurate recognition of words that were presented was also reported to be spectacularly higher in language-congruent than incongruent
conditions (Sahlin, Harding, & Seamon, 2005). At the same time, the authors interpret the appearance of false memories in cross-lingual situations as evidence for (access to) a language-independent conceptual level.

2.2.5.3 Listening comprehension

Listening comprehension research also reports evidence of language-specificity. Marian and Fausey (2006) conducted an experiment in which bilinguals were taught domain-specific information in L1 or L2. Retrieval was more accurate and faster when the language of retrieval was congruent with that of encoding. An effect on accuracy was found only in bilinguals that were highly proficient (balanced) in both languages, an effect on reaction times was found only for encoding in L1 (Spanish). In other words, when information is orally presented in the weaker language, participants do not retrieve information faster in L2 than in L1. The authors interpret these results as evidence for language-dependent memory. On the one hand, it seems that the effects can indeed be explained by participants translating the information to the target language, which is easier from L2 to L1 than from L1 to L2. On the other hand, worse L2 production compared to L1 production can account for these effects as well, since it was not harder to retrieve information in L1 when it was offered in L2. Interestingly, these results seem to suggest that memory is more susceptible to language-dependency in subjects with a higher proficiency level rather than a low level.

To conclude: bilingual memory research has not been able to settle the debate on language-dependency in memory. In addition, language can serve as a contextual clue, but there is no clarity about language-dependency in semantic memory, especially for texts. Moreover, if information is encoded in L2, this raises the question whether any semantic information is ‘lost’ before or during retrieval compared to L1.
2.3 The relation between reading comprehension and memory

Memory and reading comprehension are intertwined in several ways. Comprehension can only occur with the involvement of long-term memory, which is crucial for the understanding of words and the integration of prior knowledge, and working memory (Oakhill et al., 2014). Working memory is defined as a system in which (new or retrievable) information that is necessary to perform complex tasks is processed and maintained (Walter, 2004). The other way around, reading comprehension is a prerequisite for text memory. If a reader has difficulty with the initial understanding of a text, he/she will not be able to retain the information.

2.3.1 The reading comprehension process

Reading comprehension is a complex combination of cognitive processes (Keenan, Betjemann, & Olson, 2008, p. 294). Several theories of reading comprehension try to differentiate between these partial processes. One such view, the “Simple view of Reading”, originally described by Gough and Tunmer (1986, as discussed by Oakhill et al., 2014) states that reading comprehension is the sum of decoding ability and language comprehension. The first term refers to the ability to “decipher” the orthography of words, namely to map orthography to a lexical representation. The reader needs to understand the words in the text in the first place, since these are the building blocks of the meaning of the text. Scarborough (2001, as discussed by Oakhill et al., 2014) includes letter-sound knowledge, accurate word decoding and automaticity in decoding as the three core elements of this aspect. The latter term, language comprehension, refers to higher processes necessary to understand meaning from text or spoken word. This process is sometimes measured with listening comprehension tests: a listener does not need orthographic decoding skills to process speech, but he does need knowledge of syntax etc. Children understand spoken language rather well before they learn how to read. So language comprehension is acquired naturally during childhood, while children need instruction to learn to decode (Oakhill et al., 2014). Since the same inference-making skills and language comprehension skills are necessary for listening
comprehension as for reading comprehension, listening comprehension is a good predictor of reading comprehension (e.g. O’Reilly, Weeks, Sabatini, Halderman, & Steinberg, 2014; Kim, 2014). In reading, once the words are recognised, the reader needs to access word meanings and understand sentence and text structure to make sense of what he/she is reading. In addition, “there is more to understanding a text than the understanding of its sentences” (Cirilo & Foss, 1980, p. 108). To fully grasp what is meant, the reader usually needs to make inferences from the text, integrate background knowledge, and monitor his/her comprehension to adjust his/her inferences when necessary (Oakhill et al., 2014). When these processes all come together, a reader builds a so-called “mental model” of the text, in which the content is represented and missing links are added to create an imaginary world or situation (Oakhill et al., 2014). It is this mental model which is of interest to the current study of bilingual memory: is the conceptual model entirely language-independent or not, is it impaired after reading in L2, and how much of it is retained in the long term?

2.3.2 Factors Predicting Reading Comprehension

Different reading comprehension tests appear to measure different partial processes (O’Reilly et al., 2014), so which factors predict reading comprehension? Other models that the Simple View of Reading elaborated on the contributing components of reading comprehension. One such model is the Direct and Inferential Mediation Model (DIME) of Cromley and Azevedo (2007), which illustrates the most important predictors (background knowledge, vocabulary, strategies, and inference making) and the pathways between those elements. This model predicted 66% of the variance in reading comprehension. When it was tested for reading comprehension of scientific text, it showed a good fit, which further improved by adding a connection from vocabulary to strategies (Cromley, Snyder-Hogan, & Luciw-Dubas, 2010; see Figure 4). Still, corroborating evidence has been found for several predictors of reading comprehension that are not included in the DIME-model, such as gender (Clinton et al., 2014) and age (Ehrlich & Suez-Poy, 1995); reading motivation (Andreassen & Bråten, 2009); cognitive skills such as working memory capacity and attentional control (Conners, 2008; McVay & Kane, 2012), IQ (Keenan & Meenan, 2014); language skills such as language proficiency (in L2 research) (Droop & Verhoeven, 2003), reading fluency (Başaran,
and additional evidence was reported for some of the factors that are present in the model (such as vocabulary knowledge and prior knowledge; Coiro, 2011; Mehrpour & Rahimi, 2010). While these factors are all inherent to the reader of a text, some aspects of the text itself also affect comprehension. For example, the position of a proposition in a text will affect (to a certain extent) how well it is remembered: recency and primacy had small but stable effects on free recall in a study of Freebody and Anderson (1981). Rated importance of propositions also predicted performance in that study. Furthermore, the better the structure of a narrative text corresponds to fixed story structures known to readers, the better their comprehension will be (Thorndyke, 1977). So in narrative texts, readers map story-specific information unto a general story framework to facilitate comprehension. Though readers initially might not have such a framework available for scientific texts, evidence has shown that training providing such a framework improves recall. For example, structure strategy training facilitates recall of expository prose (Raymond, 1993) and of scientific texts (J. N. Davis, Lange, & Samuels, 1988) in a second language. When assessing memory for texts, as many factors that predict reading comprehension as possible should be taken into account.

---

**Figure 4.** Updated version of the path diagram for the Direct and Inferential Mediation (DIME) model (originally by Cromley & Azevedo, 2007, update by Cromley, Snyder-Hogan, & Luciw-Dubas, 2010)
2.3.3 Reading comprehension in L2

While all these factors predict reading comprehension in L1, some factors might be more predictive of L2 reading than others. A meta-analysis showed that L2 grammar knowledge, L2 vocabulary knowledge and L2 decoding were the strongest predictors (Jeon & Yamashita, 2014), which indicates that the lower-order processes are more important in L2 than in L1. The authors expected these three factors to have a large influence because they are most frequently reported by L2 reading researchers, together with L1 comprehension. Though L1 comprehension was also a significant correlate, it was of less importance than L2-related predictors (the ones mentioned above and others such as L2 phonological decoding etc.). These findings suggest that, during the processing of L2 text, more attention will be directed to the lexical level, since these superficial processes are the strongest predictors of comprehension outcome. Still, some authors argue that reduced comprehension in L2 is caused by difficult access to reading comprehension skills that readers have acquired in L1, rather than by issues with decoding processes or to a transfer of those skills (Walter, 2007). More specifically, Walter (2004) argued that what goes wrong in L2 reading comprehension at lower proficiency levels is the building of a well-structured mental model, possibly as a consequence of a less developed verbal working memory in L2. This view can be explained by the ‘threshold hypothesis’ which states that a lack of language-specific knowledge (below a certain threshold) cannot be compensated by (metacognitive) knowledge of reading strategies, goals and text characteristics. In other words, the low proficiency in L2 “short-circuits the transfer of reading skills” to a foreign language (Schoonen, Hulstijn, & Bossers, 1998, p. 72). In this respect, it is interesting that Oded and Walters (2001) succeeded in manipulating the quality of the mental model in L2 reading. When subjects had to write a summary of a text, they scored higher on a comprehension test afterwards, then when they had to list details from the text. This shows that, in L2, the level at which attention is directed influences the creation of a mental model. This is in accordance with the LEVELS-OF-PROCESSING (LOP) effect (Craik & Lockhart, 1972) which states that deeper processing results in a more elaborate and longer lasting memory trace. In other words, in L1, the same manipulation would probably also lead to a better mental model. Still, this LOP-effect is smaller in L2 than in L1, at least in a word list memory study (Francis & Gutiérrez, 2012).
2.3.4 Pitfalls in reading comprehension assessment

A possible approach in studying memory for texts is to select a widely used reading comprehension assessment measure and translate it. Unfortunately, many such tests are poorly validated or have been proven unreliable. Keenan et al. (2008) proved that different standardized tests are predicted by different factors, which means they are measuring different component processes of comprehension (as mentioned in the previous section). Furthermore, Coleman, Lindstrom, Nelson, Lindstrom, and Gregg (2010) detected an essential problem with the Nelson-Denny Reading Test, a test widely used in the US, recommended as a first measure of reading (dis)ability by the Law School Admission Council (LSAC). The authors presented the test to a group of participants that had not read the passages. The participants scored significantly above chance, which means that the questions can be answered based on background knowledge (Coleman et al., 2010). As a consequence, the reading skills of students with reading disabilities but with sufficient background knowledge will be overestimated by the test and they will not be offered accommodations. In the same way, if test scores are compared in L1 and L2, but questions can be answered based on background knowledge, the test might not measure language-dependent memory but the availability of background knowledge. To avoid such issues, researchers should check for “passageless” comprehension of their tests by testing the performance of subjects who have not read the passage before the test. Note that the risk of “passage independence” of questions is supposedly higher in multiple choice tasks, because (in addition to background knowledge), verbal reasoning can be used to exclude some options and certain distractors may guide readers to the right answer (Andreassen & Bråten, 2009).

Even if these validity issues were solved, the selection of an appropriate test should consider the age or participant group it was developed for, whether or not the norms are still up-to-date, and what kind of reading is investigated (reading to learn versus reading to understand, for example).

Furthermore, different test types (free recall, open questions, multiple choice, true/false judgement, or cloze tests) might induce different strategies and outcomes. This might be one of the reasons that different reading comprehension tests tap into different component processes (as explained earlier): They seem to use different test types
arbitrarily. In free recall, language production plays a crucial role, while in true/false questions, it does not. In multiple choice questions, the process of choosing a particular alternative involves reading comprehension during the test in addition to during reading the text (someone might have understood and remembered the text rather well, but might use the wrong reasoning in selecting the response alternative and get a lower score). In conclusion: selecting or creating a test to measure bilingual memory involves considering different predictors of reading comprehension, aspects of comprehension, text types, test types, and differences between languages. For this reason, we chose to develop our own materials for the question we wanted to address: how does L2 influence memory for university course materials? In the current dissertation, several text and test types are used.

2.3.5 Text accessibility

The terminological use of reading comprehension in the literature is rather ambiguous. Some authors define reading comprehension as the understanding of a text, independent of text accessibility. Text accessibility refers to whether subjects can look back at the text during the test or not. Other authors state that readers always must have access to the text in a test of reading comprehension. So retrieving information from a text that is no longer available to the reader is often viewed as reading comprehension, while we consider this to be recall.

Many studies have thus used the recall task as a measure for reading comprehension, while, as Chang describes, the reader’s task not only involves “to comprehend the text but also to recall the information accurately” (2011, p. 904). This practice of using the recall task has been subject to research in itself. Lee (1986) describes a series of studies using the free written recall task to measure L2 reading comprehension and points out the methodological difference between those studies concerning the language of recall and the pre-reading instructions announcing a recall task. Since the recall task has been used to measure L2 reading comprehension, there are some studies available on what we consider memory for texts in a second language that can easily be overlooked. These will be taken into consideration in Chapter 4.

Still, comprehension and recall are not entirely connected, as was reported by Ehrlich and Suez-Poy (1995), who found that better comprehension leads to better recall
in young subjects, but not as much in older subjects. In addition, the fact whether the
text is accessible or not has shown to change the prediction size of factors such as
reasoning ability (Andreassen & Bråten, 2009; Schaffner & Schiefel, 2013). In other
words, reading comprehension with text accessibility and recall afterwards are in some
ways different processes. Therefore, we take reading comprehension and its factors into
account, but we will only make conclusions about memory.

3. The current dissertation

The general aim of this dissertation is to explore whether learning from texts in a
second language results in worse performance than in a native language, with English as
the second language throughout the whole dissertation. Still, ‘performance’ or ‘memory’
can be operationalised in many ways. The following paragraphs provide some
background for the general methodological choices and the research lines that drove the
empirical research.

3.1 Some methodological considerations

3.1.1 Comparing languages

To find out whether studying texts in a second language results in worse memory
than in a first and to what extent, performance is directly compared for a very
homogenous participant pool in L1 and L2. This means that the information offered in
the text ought to be exactly the same, and that other textual factors such as difficulty
ought to be as similar as possible in both languages. Textual complexity is topic of a large
body of research, and several tools have been developed to assess the complexity or
“readability” of texts (Bailin & Grafstein, 2001; De Clercq et al., 2014). Still, to our
knowledge, no tools exist to compare complexity between languages, because complexity is inherently different in different languages.

To make the textual factors as similar as possible for our first set of materials, we therefore opted for matching the words in the text on frequency and prevalence (see Chapter 1) as closely as possible, choosing other translation equivalents if large differences were found. In addition, counterbalancing the texts across language conditions also avoided confounds if any substantial differences would have arisen during translation. To our knowledge, previous research has not taken these measures into account for memory tests.

### 3.1.2 Power

In all the empirical chapters, our aim was to test a sufficiently large sample to detect an effect size of Cohen’s $d = .4$. This was based on Ferguson’s claim that an effect of this size is relevant to applied settings such as education (2009). Many older studies investigating bilingual memory were severely underpowered (as discussed in previous paragraphs), and in recent years, the field of experimental psychology has also unveiled issues like the publication bias and a so-called replication crisis (G. Francis, 2012; Fritz, Scherndl, & Ku, 2014; Simmons, Nelson, & Simonsohn, 2011; Stroebe & Strack, 2014). Consistently testing with a predefined sample size and calculating effect sizes reduces the risk of drawing unsolid conclusions. As a consequence, the number of independent variables was limited in every chapter as well, so that an answer could be formulated on a limited number of questions, but with reasonable certainty.

### 3.1.3 Excluding other factors

To avoid confounds from other than language-related factors, homogenous groups were tested on several control variables (so that group differences could be checked in between-subject designs and compared between chapters). Furthermore, at the time our materials were developed, it was unclear whether reading from screen differs from reading from paper, though more and more studies on the topic are emerging (e.g. Mangen, Walgermo, & Brønnick, 2013; Myrberg & Wiberg, 2015; Porion, Aparicio, Megalakaki, Robert, & Baccino, 2016; van den Broek, Segers, & Verhoeven, 2014).
handbooks are still used in higher education, we opted to provide all texts on print. Lastly, a time limit was used in all experiments. This means subjects could not compensate for a disadvantage in L2 by spending more time on the task. It is of course possible that the effects reported here disappear when a reader can take plenty of time to cope with the more challenging context of L2.

3.2 Research lines and chapters

Since our main research question is whether memory for texts is impaired in a second language, Chapter 2 starts with a large-scale study comparing memory in an L1-to-L1 versus an L2-to-L2 condition. Since memory can be operationalised in many ways, we decided to include two test types tapping into two different aspects of memory: recall on the one hand, which can be cued or uncued, and recognition on the other hand. This way, we could find out which measure was most sensitive and whether language effects were influenced by test type. Recall was operationalised by a free recall task, in which participants write down as much as they can remember in their own words. Recognition was operationalised by a true/false judgement test, in which participants judge whether a proposition was present in the text or not. Multiple choice questions, which could be categorised as a recognition task as well, tap into comprehension processes during the task itself and were therefore deemed less adequate to test mere recognition. Another way to conceive the difference between recognition and recall tasks is as testing ‘marginal knowledge’ (Berger, Hall, & Bahrick, 1999), knowledge in memory that cannot be retrieved without the help of memory cues, versus knowledge that can be actively retrieved. Texts were taken from a test-enhanced learning study (Roediger & Karpicke, 2006) in which free recall is tested. Translations were matched on linguistic variables, and true/false tests were developed for the specific purpose of this study. Since we had two texts at our disposal, a mixed design was used, with two language groups of 100 participants each and test type as a repeated measure. To our surprise, an L2 disadvantage was found only in the recall condition with an effect size of Cohen’s $d = .8$ (large). This led to two possible interpretations: either a production deficit is the cause of this disadvantage, or the mental model of the text is less detailed with weaker memory traces, which does suffice for a recognition task.
If memory traces are weaker, they decay more easily. Hence, forgetting ought to appear more on recognition tests in a second language than in a first. In Chapter 3, an experiment (N = 172) was set up with the same materials, in which participants’ recognition memory was retested after a day, a week, or a month. Apart from clarifying the state of the mental model, this provides an insight in the forgetting curves in L1 and L2. In this case, language was tested as a within-subject variable and groups were used for the time of the second test (it was impossible to test subjects up to four times without too much interference of the testing effect). The forgetting curves were highly similar in L1 and L2 and effects were only found for the time of the test. This result is reassuring for the use of English as a medium of instruction and at the same time goes against the hypothesis that the memory traces are weaker in L2.

Since 172 subjects were present for two lab sessions for this experiment, more tests were administered in this group (for N after excluding invalid data, see Table 1). Chapter 4 reports on two experiments exploring the same research question (is memory for texts hampered in L2?) from a broader scope. Different texts, different test types, and cross-lingual conditions were selected. In Experiments 1A and 1B, handbook texts in psychology (closer to the subjects’ domain and an example of the texts they have to cope with in their studies) were followed by a test with open questions. Language was treated as a within-subject variable, comparing L1-L1 and L2-L1 conditions in experiment A, and L2-L1 and L2-L2 conditions in experiment B. No significant effects were found, though the trend was in the direction of our other findings, namely L1-L1 > L2-L1 > L2-L2. In Experiment 2, a fragment of an actual academic article in educational sciences was selected, based on the work of De Bruyne and Valcke (in preparation). Only one article was included, to keep the experimental procedure doable, and thus, only two language conditions were included as a between-participant variable. All subjects received a Dutch test, but half received the text in English, half in Dutch. The memory test included both open questions and multiple choice questions, and a language effect was found for both. Since the time pressure was higher in this experiment than in others, we concluded that factors such as time and difficulty are involved in the fact whether an L2 disadvantage occurs, which might explain inconsistencies in the literature as well. As a consequence, theoretical accounts assuming a disadvantage during the encoding process were reconsidered.
Finally, our aim in **Chapter 5** was to disentangle the large recall cost that had been found in the first chapter. As these findings are consistent with both an encoding and a production explanation we decided to adjudicate between these explanations by directly comparing unilingual and cross-lingual conditions. Language of text was a repeated measure in Experiment 1 (N = 56) and language of recall was a repeated measure in Experiment 2 (N = 59). A significant language effect was only observed for the first experiment (d = .6), which means the language of encoding is crucial for the initially reported recall cost. It is important to note that a production effect might be found in experiments in which language mistakes are punished, which we avoided as much as possible. These results have the theoretical implication that deep encoding in L2 – at least under time pressure – is problematic. They can be accounted for by language non-selective access frameworks, which will be discussed in depth in the final chapter of this dissertation. Table 1 gives an overview of the experimental set-up of every chapter investigating memory for texts.

In addition, some exploratory analyses on the data from the other chapters are presented in **Chapter 6**. The analyses revolve around two topics. First, it is described how students report on their own motivation and attitudes towards English as a Medium of Instruction. In general, they report positive feelings. Second, we have explored which of our control variables predict memory performance in Chapter 2. Results emphasize the importance of working memory in L2, though the data sets are too small to draw strong conclusions. Chapter 7 concludes with the general discussion, in which the empirical findings are tied together and both theoretical and practical implications are discussed.
<table>
<thead>
<tr>
<th>Section/experiment</th>
<th>Research question</th>
<th>Findings</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2</td>
<td>Is memory for texts in L2 weaker than in L1 for recognition or recall?</td>
<td>Large L2 recall cost</td>
<td>No recognition cost</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Is the forgetting curve steeper in L2 than in L1?</td>
<td>Similar forgetting curves (L1-L2 &gt; L1-L1 trend)</td>
<td>No significant effects (L1-L1 vs L2-L2 open questions)</td>
</tr>
</tbody>
</table>

**Table 1. Overview of the empirical chapters reporting on memory for texts in L2.**

<table>
<thead>
<tr>
<th>Chapter 4 – exp. 1A</th>
<th>Expository texts The Sun, Sea Otters (Roediger &amp; Karpicke, 2006)</th>
<th>53</th>
<th>63</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 4 – exp. 1B</td>
<td>Expository handbook texts on subliminal perception (Bryzaert, 2016; Bryzaert &amp; Rastie, 2006)</td>
<td>124</td>
<td></td>
</tr>
<tr>
<td>Chapter 4 – exp. 2</td>
<td>Fragment from scientific article on metacognition (Mayer, 1998, as translated and find in cross-lingual tasks)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 5 – exp. 1</td>
<td>Expository texts The Sun, Sea Otters (Roediger &amp; Karpicke, 2006)</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>Chapter 5 – exp. 2</td>
<td>Expository handbook texts on subliminal perception (Bryzaert, 2016; Bryzaert &amp; Rastie, 2006)</td>
<td>59</td>
<td></td>
</tr>
</tbody>
</table>

What is the contribution of L2-L1 vs L1-L1 recall + true/false L2 encoding cost, L2-L1 vs L2-L1 recall + true/false L2 encoding cost, L2-L1 vs L2-L1 recall + true/false L2 encoding cost in the free recall cost?
References


INTRODUCTION


STUDYING TEXTS IN A SECOND LANGUAGE:
THE IMPORTANCE OF TEST TYPE

Heleen Vander Beken & Marc Brysbaert

https://doi.org/10.1017/S1366728917000189
Little is known about the extent to which information encoding and retrieval differ between materials studied in first and second language (L1 and L2). In this study we compared memory for short, expository texts in L1 and L2, tested with a free recall test and a true/false judgement test. Our results show that students performed at the same level on the recognition test in both languages but not on the free recall test, with much lower performance in L2 than in L1, defined here as the dominant language. The L2 recall cost suggests that students’ performance may be underestimated if they are exclusively tested with essay-type exams in L2.

Keywords: bilingualism, text memory, language of thought
Introduction

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

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

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

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

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

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

Against the view of language-independent thought representations, there is some evidence that thought representations may not be completely language-independent (also see Alba & Hasher, 1983, for a review of the evidence that discourse memories may include more surface details than assumed by theories based on language-independent representations). First, autobiographical memory seems to be partially language-dependent. Memories of events are explained in more detail in the language in which the event took place and tend to differ depending on the language of the memory cues provided (Marian & Fausey, 2006; Matsumoto & Stanny, 2006; Schrauf et al., 1998). Second, Watkins and Peynircioglu (1983) presented their participants with mixed lists of eight Spanish and eight English words. At the end of each list, participants were given word fragments, which they had to complete. Some of these word fragments were from a target word presented (e.g., -lo--for cloud), others were from the translation of a word presented (e.g., -lo-- for the Spanish equivalent nube). If participants were unable to provide the correct word, more letters were added to the fragment (-lo-d; clo-d) until the participant was able to give the correct response. Watkins and Peynircioglu (1983) observed that more letter cues had to be given when the word fragments referred to translations than to the target words themselves, suggesting that the information stored in memory included more than language-independent semantic representations. A third
piece of evidence was reported by Marian and Fausey (2006), who ran an experiment in which bilinguals were taught domain-specific information from auditory input in L1 or L2. Retrieval was more accurate and faster when the language of retrieval was the same as that of encoding, at least for highly-proficient bilinguals. Finally, multilinguals prefer to do arithmetic in the language used at school. Apparently, counting and tables of multiplication and addition are encoded in a language-specific way (Van Rinsveld, Brunner, Landerl, Schiltz, & Ugen, 2015).

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

To conclude: Psycholinguists thus far have done little systematic research on encoding and retrieving text information presented in L1 or L2 (see the discussion section for two small-scale studies). In their models of word processing, they assume the existence of language-independent meaning representations, to which the language input must be translated and which are translated again for verbal output, in line with ideas developed in the 1960-1980s. At the same time, there is some evidence that memories for text and discourse may be more accessible in the language studied than in another language mastered.

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

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

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

Finally, the way in which memory is tested is likely to make a difference as well. Traditionally, a distinction is made between recognition and recall (e.g., Gillund & Shiffrin, 1984). Though both test types tap into declarative memory (Haist, Shimamura, & Squire, 1992), the processes of retrieval and the conditions for success differ (Hogan & Kintsch, 1971). Recall involves an extended search, which is slow and uncertain and which requires more processing resources, as is proven by a decreased recall performance (compared to recognition) with increasing age (Craik & Mcdowd, 1987). A recognition test includes many more cues, so that memory traces can be accessed more directly. In particular, true/false judgements can be considered as “locating questions” according to Guthrie (1998, see also Tal et al., 1994). Guthrie pointed out that the processes needed to locate details in a text are distinct from the processes involved in recalling the main ideas of the same text. Similarly, it may make a difference if one has to match a detailed (“locating”) question to stored information than when one has to reproduce the core ideas from that same memory without cue. Alba and Hasher (1983) provided evidence that recall tests are more influenced by the participant’s memory schemas and scripts than recognition.

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

When confronted with such a multitude of potentially important variables, it is tempting to run a series of small experiments, addressing the various questions and possible confounds. A danger in doing so, however, is that each experiment tends to be underpowered, because of resource constraints. As has been well documented, this
involves two risks. The first is that a null effect is obtained, which cannot be interpreted. The second is that a significant effect is found, which cannot be replicated (Gelman & Carlin, 2014), in particular when effects are close to the significance level (Francis, 2012; Leggett, Thomas, Loetscher, & Nicholls, 2013; Simmons, Nelson, & Simonsohn, 2011). To avoid these problems, we ran a power analysis before setting up the experiment (see under Method).

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

Method

Participants

To decide on the number of participants needed for a sufficiently powered experiment, we started from the observation that an effect size of $d = .4$ is seen as a practically significant effect. Such minimum effect size is usually required for efficient therapies and for group differences that must be addressed in applied settings (e.g., education; Ferguson, 2009). Since our design included a between-groups variable, we needed two groups of 100 participants to have 80% chance of detecting an effect of $d = .4$ (Cohen, 1992; see also Callens, Tops, & Brysbaert, 2012).
A total of 199 first year psychology students from Ghent University took part in the experiment in partial fulfilment of course requirements and for an additional financial reward (data collection was planned for 200 participants, but one student did not show up on any of the sessions they were invited to). All participants were Dutch native speakers who had studied English in high school for at least four years and who were regularly exposed to (subtitled) English television programs and English songs. In some of their university courses English handbooks were used, even though the teaching happened in Dutch. Note that, in this study, L1 was defined in terms of dominant language, not as the first acquired language. The data of four students who did not have Dutch as their dominant language were excluded from all analyses, so that the final analyses are based on N = 195. The participants’ mean age was 18.6 yrs (sd 2.3); 129 were female students, 66 male. Participants were randomly assigned to the conditions.

Materials

Texts

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

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

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

*Free recall and true/false judgement tests.*

Two types of tests were administered to accompany the texts: a free recall test and a true/false judgement test. In the free recall test, participants received the following instruction: “Write a summary of the text you have just read. Be as detailed as you can be”. This way, participants were not asked to literally reproduce the text, but to produce the ideas and we encouraged them to add details when possible.

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

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

*Motivation and Text-related Questionnaires.*

After the true/false and free recall tests, the participants completed two questionnaires. The first asked about their general attitude towards reading and testing: their testing motivation, their self-perceived level of performance relative to fellow students, and their general reading motivation in Dutch (L1) and English (L2). The second questionnaire checked for prior knowledge about the texts, the perceived difficulties (both content and structure) of the texts, and how interesting the texts were. The questionnaires were presented in Dutch to all participants, using 7-point Likert scales. See Appendix D.1 for the questionnaire.

*Subjective assessment of language proficiency.*

The participants’ language background was assessed with a Dutch version of the Language Experience and Proficiency Questionnaire (Marian, Blumenfeld, & Kaushanskaya, 2007; translated by Lisa Vandeberg; adaptation Freya De Keyser, Ghent University, and Marilyn Hall, Northwestern University).

*Objective L1 proficiency tests*

L1 proficiency was measured with the Dutch LexTALE test, a language-specific lexical decision test containing 40 words of various difficulty levels and 20 nonwords (Lemhöfer & Broersma, 2012). In addition, the participants received a semantic vocabulary test in a multiple choice format with four answer alternatives and a Dutch spelling test (see Appendix C.2) in which they had to spell words of various spelling difficulties that were read aloud (all developed at the department).

*Objective L2 proficiency tests*

L2 proficiency was measured with the English LexTALE test of vocabulary knowledge for advanced learners of English (Lemhöfer & Broersma, 2012). Next, the participants received a version of the MINT picture-naming task, adapted for Dutch-English speakers (Gollan, Weissberger, Runnqvist, Montoya, & Cera, 2012), in which they saw a black- and white picture of an object of which they had to type the English
name. The Oxford Quick Placement Test (QPT; 2001) was also administered, which is considered a measure of general proficiency consisting of multiple choice items of vocabulary and sentence comprehension and grammar (verb use, part of speech regulations, ...). Finally, an English spelling test was given, similar to the Dutch spelling test (see Appendix C.2).

Measures of reading exposure, intelligence and WM

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

Distractor task between learning and testing

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

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

<table>
<thead>
<tr>
<th>Study phase I</th>
<th>Dutch</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test phase I</td>
<td>De zon</td>
<td>Zeeotters</td>
</tr>
<tr>
<td>Test phase II</td>
<td>Zeeotters</td>
<td>De zon</td>
</tr>
</tbody>
</table>

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

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

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

Results

Scoring

In our marking of the free recall tests we followed the guidelines set out by Roediger and Karpicke (2006). We scored the presence and correctness of the ideas from the text, irrespective of spelling errors and the overall organization of the recall protocol. Participants received 1 point for every correctly recalled idea and 0 points if the idea was recalled incorrectly or not recalled at all. If an idea was partially recalled, a .5 score was given. For the text about sea otters, three propositions had to be split into two separate ideas, because often only one of them was recalled. A random sample of 100 recall forms was scored by two raters: the first author and a Dutch-English teacher with test rating experience. The second rater got the following guidelines: Spelling and grammatical mistakes must not be punished unless those mistakes obscure meaning (a similar guideline is given for the PISA tests; see appendix in Cartwright, 2012). We first calculated the interrater reliability: the Pearson correlation between the scores of both raters across all forms was .85. Partial analyses showed similar results of $r = .83$ for the Dutch ratings only, $r = .87$ for the English texts, $r = .83$ for The Sun and $r = .86$ for Sea Otters. Given the reassuring correlation, the rest of the tests were rated by the experimenter only. Since both raters barely ever used the .5-score (33/3210 trials) and did not agree with each other on those scores, we replaced them by 0.

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

All data are available at https://osf.io/2twzd/ (Open Science Framework).
**Testing whether the students were matched in the L1 and L2 condition**

Because the main comparison involves L1 vs L2 studying, we first checked whether both groups were matched on the control variables we assessed. Table 1 and 2 show that this was the case. There were no significant differences between the two groups if a correction for multiple testing was taken into account\(^1\). In addition to this group comparison, we looked into within-subject differences in motivation of the total group of participants. Interestingly, participants in general had a higher reading motivation in L1 (\(M = 5.18, \text{SD} = 1.41\)) than in L2 (\(M = 4.51, \text{SD} = 1.47\); Wilcoxon signed rank test resulted in \(V = 8610.5, p < .001\)). The reliability of the measures was measured using Cronbach’s alpha, which was generally high. Only the Raven’s matrices resulted in an alpha of .47, probably due to an error in the administration (we presented each question on a central screen for the same duration, in group, while normally, the test is taken individually and participants can move through the items at their own pace). Table 3 displays the reliability measures and the correlations between the various measures.

\(^1\) All test statistics (Welch two-sample t-tests and Wilcoxon rank sum tests) for the group comparison can be found at [https://osf.io/2twzd/](https://osf.io/2twzd/).
Table 1. *Mean scores of the language groups on the various proficiency and intelligence tests (standard deviations between brackets).*

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

Note: The test statistics can be found at [https://osf.io/2twzd/](https://osf.io/2twzd/)
Table 2. Mean scores of the language groups on the self-ratings included in the questionnaire (standard deviations between brackets).

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

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

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

<table>
<thead>
<tr>
<th>Tests</th>
<th>Dutch LexTALE</th>
<th>Dutch voc. MC</th>
<th>Dutch spelling</th>
<th>Eng. LexTALE</th>
<th>Eng. spelling</th>
<th>MINT</th>
<th>QPT</th>
<th>Author recogn.</th>
<th>Raven</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch LexTALE</td>
<td>0.63</td>
<td>0.19</td>
<td>0.33</td>
<td>0.34</td>
<td>0.2</td>
<td>0.21</td>
<td>0.19</td>
<td>0.14</td>
<td>0.04</td>
</tr>
<tr>
<td>Dutch voc. MC</td>
<td>0.29</td>
<td>0.66</td>
<td>0.25</td>
<td>0.27</td>
<td>0.30</td>
<td>0.30</td>
<td>0.36</td>
<td>0.34</td>
<td>0.18</td>
</tr>
<tr>
<td>Dutch spelling</td>
<td>0.46</td>
<td>0.33</td>
<td>0.87</td>
<td>0.28</td>
<td>0.61</td>
<td>0.26</td>
<td>0.32</td>
<td>0.25</td>
<td>0.14</td>
</tr>
<tr>
<td>Eng. LexTALE</td>
<td>0.44</td>
<td>0.35</td>
<td>0.32</td>
<td>0.90</td>
<td>0.54</td>
<td>0.59</td>
<td>0.57</td>
<td>0.28</td>
<td>0.21</td>
</tr>
<tr>
<td>Eng. spelling</td>
<td>0.26</td>
<td>0.38</td>
<td>0.68</td>
<td>0.59</td>
<td>0.93</td>
<td>0.56</td>
<td>0.55</td>
<td>0.27</td>
<td>0.15</td>
</tr>
<tr>
<td>MINT</td>
<td>0.18</td>
<td>0.38</td>
<td>0.29</td>
<td>0.64</td>
<td>0.60</td>
<td>0.93</td>
<td>0.63</td>
<td>0.16</td>
<td>0.20</td>
</tr>
<tr>
<td>QPT</td>
<td>0.29</td>
<td>0.49</td>
<td>0.38</td>
<td>0.66</td>
<td>0.63</td>
<td>0.72</td>
<td>0.83</td>
<td>0.24</td>
<td>0.14</td>
</tr>
<tr>
<td>Author recogn.</td>
<td>0.05</td>
<td>0.42</td>
<td>0.27</td>
<td>0.30</td>
<td>0.28</td>
<td>0.17</td>
<td>0.27</td>
<td>0.97</td>
<td>0.06</td>
</tr>
<tr>
<td>Raven</td>
<td>0.35</td>
<td>0.33</td>
<td>0.22</td>
<td>0.33</td>
<td>0.23</td>
<td>0.31</td>
<td>0.24</td>
<td>0.09</td>
<td>0.46</td>
</tr>
</tbody>
</table>
Assessing the participants’ L2 proficiency level

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

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

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

Performance on the memory tests

To analyse memory performance, we used a 2 (language group) x 2 (test type) mixed ANOVA. Participants had been divided in eight groups, each presented with only one language and one combination of text and test type in a counterbalanced order. Given that the texts and presentation orders were control variables, to be counterbalanced, they were not included in the analysis. As each participant did one recognition and one recall test, this was a repeated measure. The analysis indicated a significant main effect of

---

In the self-ratings participants indicated they had more prior knowledge about The Sun (M = 3.3 on a 7-point rating scale) than about Sea Otters (M = 1.8) but that the text about the Sun was experienced as more difficult than the text about sea otters (3.6 vs. 3.1). Both texts were judged to be matched in terms of structural difficulty (M_{The Sun} = 2.89 and M_{Sea Otters} = 3.03) and in terms of power to interest (M_{The Sun} = 4.6 and M_{Sea Otters} = 4.5).
language (F(1,193) = 19.88, p < .001, η²_p = .09), a significant main effect of test type (F(1,193) = 286.79, p < .001, η²_p = .59, Type III Anova) and, most importantly, a significant interaction between both variables (F(1,193) = 30.25, p < .001, η²_p = .14). Figure 2 shows the effects (see also Table 4 for the exact data). Separate comparisons indicated that the difference between L1 and L2 was not significant for the recognition test (Cohen’s d = .07; F(1,193) = .26, η²_p = .001)\(^3\), but resulted in a large effect size for the recall test (d = .86; F(1,193) = 35.68, p < .001, η²_p = .16). Cohen’s d values of .2 are usually considered “small”, .5 “medium” and .8 or higher as “large”. As indicated in the method section, d values of .4 and more are considered to be of practical relevance in applied settings. To check how large the evidence for the effects was, a Bayesian version of the analysis was conducted. The Bayes factor (BF\(_{10}\)) for a model with language (compared to the null model) was 7.92, which means there is moderate evidence for a language effect. The Bayes factor (BF\(_{10}\)) for a model with test type (compared to the null model) was 1.02×10\(^7\), which indicates extreme evidence for an effect of test type (the fact that there is a 50% chance level in one test contributes to this, see below). To evaluate the evidence for the interaction, the full model was compared with a model omitting the interaction (top-down analysis), resulting in BF\(_{10}\) = 7.83×10\(^{-6}\) for the omitted version, which is extreme evidence for the full model.

Since the true/false test had only two response alternatives, it had an estimated minimal performance level of 50% (chance level), hampering the comparison of the recognition test with the recall test\(^4\). A simple equation to correct for this, is to recode the obtained recognition scores with the equation corrected_score = (raw_score – chance_score) / (maximum_score – chance_score). Applied to the yes/no test, a raw score of .80 results in a corrected score of (.80 - .50) / (1.00 - .50) = .60 (or 60%). In such an analysis it is customary to level all scores under 50% to zero performance. When the analysis was redone with the corrected scores, the same pattern of results was obtained. The main effect of language remained significant (F(1,193) = 11.14, p < .01, η²_p = .05), as did the main effect of test type (F(1,193) = 7.12, p < .01, η²_p = .04, Type III Anova). Most importantly, the interaction between both variables remained significant

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

\(^4\) The authors thank an anonymous reviewer for pointing to this issue.
(F(1,193) = 12.76, p < .001, η²_p = .06). A separate comparison indicated that the difference between L1 and L2 was not significant for the recognition test (Cohen’s d = .08; F(1,193) = .29, η²_p = .001). Of course, the effect remained the same for the recall test, as this variable was not altered (d = .86; F(1,193) = 35.68, p < .001, η²_p = .16).

**Fig. 2** Mean percentage of recalled ideas in all conditions with 95% confidence intervals. Note that chance level for a true/false test equals 50%. So the average scores on this test could be compared to a 60% score for the free recall test.
Table 4. Means, standard deviations and ranges of the scores in the true/false judgement test and the recall test as function of the language in which the text was studied and the test taken.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True/false judgement</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 (N = 100)</td>
<td>80.9</td>
<td>11.8</td>
<td>46.7</td>
<td>96.7</td>
</tr>
<tr>
<td>L2 (N = 95)</td>
<td>80.1</td>
<td>8.7</td>
<td>53.3</td>
<td>96.7</td>
</tr>
<tr>
<td><strong>Free recall</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 (N = 100)</td>
<td>56.3</td>
<td>14.2</td>
<td>23.3</td>
<td>90.0</td>
</tr>
<tr>
<td>L2 (N = 95)</td>
<td>44.1</td>
<td>14.3</td>
<td>15.2</td>
<td>80.0</td>
</tr>
</tbody>
</table>

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

Discussion

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

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

The recognition test was a test to measure as much knowledge as possible, including marginal knowledge, as defined by Cantor et al. (2014). Participants were given detailed statements and asked whether those were true or false according to the text.
Against our expectations, students did not show any difference in performance on the true/false test as a function of the language in which they had studied the text. They were correct on 80% of the questions (corresponding to a 60% score if corrected for guessing), both when they had studied in L1 or in L2 (Figure 2). This suggests that students understood the study materials equally well in L2 and L1 and did not perform at ceiling level.

In contrast, participants studying in L2 performed significantly worse on the free recall test (44%) than the participants studying in L1 (56%). The difference corresponds to a large standardized effect size of $d = .86$, meaning that it is of practical relevance in applied settings. In the remainder of the text we will call this difference the L2 recall cost. Because of the large number of participants we tested and the many precautions we took to make sure that both groups were matched, we can have confidence in the reliability and the replicability of the effect.

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

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

An interesting idea in this respect is that it may be possible to train the translation of thoughts into L2 output. Karpicke and Roediger (2007) observed that their students remembered more from a text after having taken a test than after been given the opportunity to study the text for a second time. One of the explanations they proposed for this “testing effect” was that taking a test provided students with practice in retrieval

---

5 We have since replicated this effect and extended it to intervals of 1 week and 1 month, as will become clear in Chapter 3. So, the equivalent performance on the recognition test in L1 and L2 is unlikely to be due to the short time period between the study phase and the test phase.
processes. If this explanation is valid, we may be able to diminish the L2 recall cost by providing L2 students with practice in L2 recall before they take a test (or exam). In this respect, it may also be of importance that our participants did not know beforehand which test they were getting. It may be that students study differently if they know they will have to take a written essay-type exam in L2.

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

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

Two other studies are relevant in this respect. Chen and Donin (1997) asked 36 Chinese-English bilinguals to read a short biology text in either Chinese or English, using a cross-lingual design with L2-L2, L2-L1 and L1-L1 conditions. Half of the students were biology students with high background knowledge of the topic; half were engineer students with a limited background. Participants were asked at four places within each passage to orally recall what they had just read, and they were asked to give detailed recall of the whole passage at the end of each text. The quality of recall differed as a function of the background knowledge but, surprisingly, it did not differ significantly between the language conditions tested, contrary to what we found. However, the condition L2-L2 with L2 recall seems to show a trend towards lower performance than the L2-L1 condition. The study may have lacked the power to pick up the difference. In addition, participants spent more time to read the text in L2 (remember we had a fixed studying time of 7 min).
Longer reading times were also reported by Donin et al. (2004). They asked 16 Canadian army officers to read English (L1) or French (L2) texts and to retell in English what they had read after every 4 sentences and after the full text. The participants needed more time to read in L2 than in L1, but memory accuracy after reading was equivalent. Again, however, the power of the study was very low.

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

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

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

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

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

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

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

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

Although our findings raise a list of theoretical issues (some of which we hope to address in the future, and some of which we hope others will find interesting to tackle), they do point to an important practical implication. The observation that students have a serious L2 recall cost and at the same time good L2 recognition performance, raises the question of what type of test they should be given for their exams. If all exams are essay-type exams, it is to be feared that L2 students will be at a serious disadvantage to obtain good grades (unless training helps them to acquire these skills and such training is offered to the students before they have to take their exams). On the other hand, if all exams are of the recognition type, L2 students may find themselves even less able to talk about their knowledge (in L2). Much here, of course, depends on the type of skills taught in the course. If the skills are verbal, it can be defended that students should be able to express themselves in the language of their study. However, the situation becomes more complicated for less verbal skills. An L2 engineering student, for instance, may have learned perfectly how to design a machine, but not be able to explain this at the same level in L2 essay writing.

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


STUDYING TEXTS IN A SECOND LANGUAGE:

NO L2-DISADVANTAGE

IN LONG TERM RECOGNITION MEMORY

Heleen Vander Beken, Evy Woumans, & Marc Brysbaert

Despite an increase in bilingualism and the use of English as a medium of instruction, little research has been done on bilingual memory for learnt information. In a previous study, we found an L2 recall cost but equal recognition performance in L2 versus L1 when students studied short expository texts (Vander Beken & Brysbaert, in press). In this chapter, we investigate whether there is a recognition cost after a longer delay, which would indicate that the memory trace is weaker in L2. Results showed equal performance in L1 and L2, suggesting that the recall cost is either located at the production level, or that the levels-of-processing effect is mediated by language, with unaffected surface encoding leading to effective marginal knowledge on the one hand, and hampered deep encoding leading to ineffective (uncued) recall. This chapter also contains the Dutch vocabulary test we used for native speakers.

Keywords: bilingualism, learning from text, long-term memory, levels-of-processing effect
Introduction

Globalisation has led to an increasing number of people that communicate or study in another language than their native tongue. In the European Union, for example, the number of monolinguals has decreased to 46% in 2012 (TNS Opinion & Social). In addition, English is becoming more and more dominant, taking the role of a lingua franca (knowledge of some other languages is even decreasing as a consequence; TNS Opinion & Social, 2012). Despite the internationalisation of education and the increasing use of English as a medium of instruction (EMI), little research has been done on the consequences of studying in a second language. Still, with every start of a new academic year, the debate in higher education revives: is it worthwhile to present information, teach, or test students in a language that is not their native one? From an educational perspective, is studying in a second language (L2) a “desirable difficulty” (the perspective that long-term learning occurs through difficulties in learning, e.g. Metcalfe, 2011), a challenge that makes learning just hard enough, or does it obstruct learning possibilities? To answer this question, we need to understand how information is encoded in and retrieved from memory in L2, compared to the first/native language (L1).

Declarative memory is traditionally split up between episodic and semantic memory. While semantic memory contains the gist of information about the world, episodic memory contains contextual information tied to the stored event (e.g. Graves & Altarriba, 2014). Information that is processed can be transferred from episodic to semantic long-term memory, in which the contextual information is lost. Neurologically, the hippocampus is responsible for encoding of new – hence, episodic – memories (Hardt, Nader, & Nadel, 2013). Within minutes up to hours of this initial hippocampal encoding, neocortical traces are formed. These neocortical traces are the neurological equivalent of semantic memory. In other words: all declarative memory was episodic in its initial stage. Memory consolidation is considered as the reorganization of semantic memory in which hippocampal traces are no longer needed and memory is located in the neocortex only (Hardt et al., 2013). Memory decay can thus be explained by the fact that the hippocampal memory traces are removed during sleep, while the neocortical traces are too weak to remain without the hippocampal connection (though episodic memory remains stored in and retrieved from the hippocampus (Nadel & Hardt, 2011).
Bilingualism research has explored both types of memory to some extent. The principal theoretical view about bilingual semantic memory has been that meanings of words are stored at a language-independent conceptual level which is connected to all lexicons of a multilingual (e.g. the Revised Hierarchical Model, for a discussion of this model, see Brysbaert & Duyck, 2010)). Visual word recognition research has confirmed that, both at the word and sentence level, non-target language knowledge interferes with recognition of a target language (Van Assche, Duyck, & Hartsuiker, 2012). This theory accords with an idea found in older memory research (in the 1960-80s, e.g. Alba & Hasher, 1983; Schank, 1972): when people read a text, they do not remember it verbatim but they do remember the gist. The so-called deep structure of the text remains, though the surface form is lost. As a consequence, Schank (1980) concluded meaning is represented free of language.

Nevertheless, we intuitively expect a disadvantage when reading or studying in L2. Within the language non-selective view, there are three frameworks that would explain L2 disadvantages. The cross-linguistic interference hypothesis assumes that competition of L1 lexical representations interferes in L2 recognition (see Weber & Cutler, 2004 for auditory recognition). According to some authors (Levy, Mcveigh, Marful, & Anderson, 2007; others disagree: Runnqvist & Costa, 2012), this competition results in retrieval-induced forgetting: when you retrieve a concept in one language, this process will hamper retrieval in the other. If L1 representations indeed interfere with L2-recognition, one would expect that reading or studying L2 text takes more time and that the encoding process is hindered. A second account sets out from the fact that L2 is used less frequently than L1, resulting in weaker linguistic representations (Gollan, Montoya, Cera, & Sandoval, 2008). This weaker-links hypothesis directly compares L2-items to low-frequent L1-items. Since familiarity with these items is lower, this account explains that recognition memory for words is better in L2 (the words are less familiar and, as a consequence, more unique in memory; Francis & Gutiérrez, 2012) and entails that L2 semantic representations are less detailed (Finkbeiner, 2002). A third account is located at the level of working memory. The resource hypothesis expects that the cognitive load of L2 processing is higher, resulting in less working memory capacity for other processes (Francis & Gutiérrez, 2012; Sandoval, Gollan, Ferreira, & Salmon, 2010).
Despite elaborate evidence of language interference in visual word recognition, there are reasons to consider the possibility that memory for text is language-specific. The encoding-specificity principle states that more information is remembered when the context of encoding and retrieval are similar. Four lines of research provide evidence for this principle. Firstly, in autobiographical episodic memory, people recall more events or recall events in more detail when they are asked in the language in which the event took place (Marian & Neisser, 2000; Matsumoto & Stanny, 2006; Schrauf & Rubin, 1998). Secondly, in word list recall, more words are recalled in congruent language conditions, even in the weaker language (Nott & Lambert, 1968; Watkins & Peynircioglu, 1983). Thirdly, in listening comprehension, participants are also able to recall more information in the same language condition than in a cross-lingual condition (Marian & Fausey, 2006). Finally, when people read an article in silent or noisy conditions, their recall and recognition performance is better in the context-congruent conditions (Grant et al., 1998). If these results translate to other modalities or other types of context, memory for texts might be language-specific, and a lower proficiency may result in lower memory performance. For word list recall and listening comprehension, one might wonder whether these memory traces are part of episodic or semantic memory. Since memory is tested shortly after encoding in these experimental paradigms, the memory consolidation process would not have taken place yet (following Hardt, Nader and Nadel’s view, 2013). Hence, encoding specificity is possibly an effect that is limited to episodic or hippocampal memory traces.

Apart from the dichotomy between episodic and semantic memory, there are subtypes within these categories. When discussing semantic memory (or memory in general), we need to take into account what is really tested. People possess large amounts of knowledge, but not necessarily in an active way. Knowledge that cannot be retrieved spontaneously, but can be recognised or retrieved after presentation of a cue, is called “marginal knowledge” (Berger, Hall, & Bahrick, 1999; Cantor, Eslick, Marsh, Bjork, & Bjork, 2014). A recall test will thus not only test a different type of retrieval, compared to a recognition test (Gillund & Shiffrin, 1984; Haist, Shimamura, & Squire, 1992), it will estimate the amount of accessible knowledge, leaving this “marginal knowledge” untouched. In a previous study, we investigated how both recall and recognition for L1 and L2 texts differ (Vander Beken & Brysbaert, in press). A group of 199 participants
studied short expository texts about biology topics within a limited time frame. Afterwards, they received a true/false test about one text and a free recall test about the other. We found no L2 disadvantage in recognition memory, but a significant and rather large disadvantage in L2 recall. These findings indicate that initial encoding was not problematic. Otherwise, there would be a recognition cost as well. However, test performance does suffer from weaker language proficiency in certain conditions. The question is whether this disadvantage is situated merely at the production level, resulting in dissociation between what is known and what can be produced, or at the level of encoding or storage, namely in the richness of the memory trace. Craik and Lockhart’s (1972) levels-of-processing framework explains that initial encoding processes surface form, while the following stages are responsible for the extraction of meaning. So deeper processing “implies a greater degree of semantic or cognitive analysis” (Craik & Lockhart, 1972, p. 675), also called elaboration coding, which results in a more elaborate and longer lasting memory trace.

A possible reason to assume a disadvantage in L2 elaboration encoding can be derived from the Landscape Model by van den Broeck, Young, Tzeng, and Linderholm (1999). This theory assumes that a mental model of a text consists of a “landscape” of interrelated concepts (i.e. concepts of biology and text-specific propositions for the texts in our study) that is continuously updated during reading. More specifically, when a concept is activated, it entails cohort activation as well: related concepts are co-activated to a certain extent. Despite the fact that text comprehension suffices for recognising statements about the text, the mental model might be “weaker” in L2. For example, if a domain-specific word is unknown or unfamiliar to the reader, he/she might still understand the sentence or recognise whether a statement is correct, but the concept will not be activated, nor will it activate related concepts. So the semantic richness and activation of the mental model in L2 would be smaller, which is in line with the weaker-links hypothesis that was discussed earlier in this paper. The weaker “landscape”-effect may also be mediated (or enlarged) by lower motivation for reading in L2 (Vander Beken & Brysbaert, in press), since attention also plays a role in the way concepts are translated to the mental model (van den Broek et al., 1999).

If the mental model is weaker, long term memory would suffer from additional forgetting. Memory traces that are weaker, and less easily recalled, also fade out faster. In
higher education, information has to be retained for days up to months, and in other real-life situations, retrieval of important information is still relevant after years. Hence, for the current study we decided to test memory for text after longer intervals and employ both an immediate and a delayed recognition test in L1 (Dutch) and L2 (English). The choice for the recognition test was made because (1) there was no difference on the initial recognition test in the previous study, which creates the opportunity to measure additional loss only and (2) the scores on that test were high enough to measure a decrease without dropping to chance level. We do not expect an L2 recognition cost on the immediate test based on the previous study (Vander Beken & Brysbaert, in press), but there might be such a cost for delayed recall. If the recall cost in L2 is due to L2 production only, the rate of forgetting will be similar and none of the language-interference hypotheses will be confirmed, while the encoding-specificity principle could account for that result. If we find a delayed recall cost, this would suggest that there is a cost at the earlier memory processes, namely encoding (a poorer mental model) or storage, which is in line with the weaker-links hypothesis.

In addition, we test whether memory illusions are more persistent in L2 versus L1. In the context of testing memory, false memory illusions are positive responses to lures in recognition tests (often multiple choice tests). These false memories can be created by merely presenting a false statement repeatedly, increasing the possibility that the statement is later viewed as correct, but it also seems to depend on the performance on the initial test: the illusion is rarely found when the initial answer was correct (Marsh, Roediger, Bjork, & Bjork, 2007). Inspired by this finding, we tested whether illusions arose as the consequence of lures in our test. If the memory trace is weaker in L2, due to shallow processing, we expect more illusions to arise in that language.

Method

Participants

A total of 171 first year psychology students from Ghent University participated in partial fulfilment of course requirements and for an additional financial reward. All
participants were Dutch native speakers who had studied English in high-school for at least four years and who were regularly exposed to (subtitled) English television programs and English songs. In some of their university courses English handbooks were used, even though teaching took place in Dutch. The data of five students who did not have Dutch as their dominant language were excluded from all analyses. Note that, in this study, L1 was defined in terms of dominant language, not as the first acquired language (though for most students, the dominant language was also the native language). In addition, seven students were excluded from the analysis because they reported having dyslexia, and another four for other reading or learning disabilities (such as ADD). We removed one additional participant who had not filled in any of the proficiency tests. In the resulting dataset (N = 155), mean age was 19.47 yrs (sd 4.4); 118 were female students, 33 male (four did not indicate gender). Participants were randomly assigned to the conditions. This experiment was part of a bigger set-up in which we included other experiments of text memory for 146 participants. In the procedure section, we will discuss this entire set-up, although the results of the other memory tests are not discussed in this chapter. For the remaining 25 participants, these additional experiments were dropped from the set-up (resulting in an hour less of testing).

Materials

Texts

We used two short, English texts from a study of Roediger and Karpicke (2006). Each text covered a topic in the domain of natural sciences: The Sun and Sea Otters. The English texts were translated to Dutch and the texts were matched between languages on semantics and word frequencies (see Vander Beken & Brysbaert, in press). The texts were between 248 and 279 words long. 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.

True/false judgment tests.

Roediger and Karpicke (2006) divided their texts into 30 ideas or propositions that had to be reproduced. In a previous study (Vander Beken & Brysbaert, in press), we used this list as a scoring form for free recall tests and created true/false judgment tests of 46 questions. Thirty true/false questions were derived from the ideas on the free recall scoring form. For example: “The Sun today is a white dwarf star” requires a
FALSE response since the text states that “The Sun today is a yellow dwarf star”. Next to those literal questions, 10 inferential questions were written: five inferences were based on one proposition, the other five on several propositions in the text. An example of such a question is “The surface of a red giant star is hotter than that of a yellow dwarf star”. To respond to that question, the reader has to remember and integrate information about the surface temperature of two of the mentioned star types. In addition, six lure questions were created containing a statement that was not mentioned in the text but was in some way related to a concept in the text. An example of such a statement is “Sea otters live around Alaska”, while Alaska was mentioned in the text as the location of an oil spill but not described as sea otters’ necessary habitat. All questions were translated to Dutch. The instruction for the test was “Tick the correct answer box for every statement, based on the text you have just read”. In the previous study, these questions were checked for passage-dependency in a separate group of participants who did not read the texts, resulting in the exclusion of some questions that were answered better than chance level by this separate group, indicating that they test prior knowledge rather than memory of the texts.

Since it was our goal to test participants’ knowledge of the same topic on an immediate and a delayed recall test, we needed two tests for every text. To avoid test effects due to repeated items, we created parallel tests: we selected pairs of questions that were similar in topic and difficulty but did not test the same proposition from the text. For example, when we had one question about the size of sea otters and one about their weight, the first question was included in version A of the test and the second in version B. Difficulty measures were based on test scores on the 46-item version of the test in a previous study (Vander Beken & Brysbaert, in press). Only the lure questions were repeated in both tests to investigate whether these false propositions led to false memories indeed. These questions were analysed separately and not included in the general analysis. This resulted in two parallel tests of 20 questions for The Sun (of which five lure questions) and 18 questions for Sea Otters (of which six lure questions).

The tests were administered online, using LimeSurvey (an Open Source PHP web application available through the university). Participants were obliged to answer all questions; answer options were “yes”, “no”, and “I don’t know”. The latter option was
added to avoid guessing if memory loss is large (this way, chance level scores were avoided).

The texts and the tests can be found in Appendix A at the end of this book.

**Motivation and Text-related Questionnaires**

After the immediate true/false tests, the participants completed some questions about the texts, concerning prior knowledge, perceived difficulty (of both content and structure), and how interesting the texts were. Next, a general questionnaire tapped into their general attitude towards reading and testing. The questionnaire contained single questions for their testing motivation and their self-perceived level of performance relative to fellow students, and several questions about their general reading motivation in Dutch (L1) and English (L2), and their attitudes towards EMI (mostly three questions per sum score). This information can be used to get an insight on how students experience EMI, apart from how they perform. The questionnaires were presented in Dutch to all participants, using 7-point Likert scales. It can be found in Appendix D.2.

**Subjective assessment of language proficiency**

The participants’ language background information was assessed with a selection of relevant questions from the Dutch version of the Language Experience and Proficiency Questionnaire (LEAP-Q, Marian, Blumenfeld, & Kaushanskaya, 2007; translated by Lisa Vandeberg; adaptation Freya De Keyser, Ghent University, and Marilyn Hall, Northwestern University). This was used to exclude non-dominant speakers of Dutch from all analysis.

**Objective L1 proficiency tests**

L1 proficiency was measured with a 75-item Dutch vocabulary test in a multiple choice format with four answer alternatives (developed at the department and listed in Appendix C.1).

**Objective L2 proficiency tests**

L2 proficiency was measured with the English LexTALE test of vocabulary knowledge for advanced learners of English (Lemhöfer & Broersma, 2012) and Nation and Beglar’s (2007) vocabulary size test in multiple choice format. The latter was
administered on www.vocabularysize.com, on which researchers can register and set up the test with a log-in code for the participants.

*Working memory*

Working memory capacity was measured with an automated operation span task programmed in E-prime 2.0.10 (Unsworth, Heitz, Schrock, & Engle, 2005).

*Text-specific vocabulary knowledge*

The delayed English true/false tests were followed by a text-specific vocabulary test in which participants had to explain, translate or give a synonym of the more central or low-frequent words of the texts (10 words for *The Sun*, 14 for *Sea otters*). Both English and Dutch answers were considered correct.

*Procedure*

Tests were administered in groups of 33 participants at most. Every participant had to be present for two lab sessions and fill in some questionnaires at home, equivalent to three hours in total (lab and online time combined). There were interval groups of 1 day, 7 days or 30 days (plus or minus one day for the two last groups). Students registered online for the sessions, so the interval groups were created based on their availability. We had several subgroups for all three interval groups and selected different times of the day and week to avoid effects of fatigue.

All participants received one text in English and the other in Dutch. The language-text relation was counterbalanced across subjects. All tests were presented in the language of the text. Since there were two parallel tests for every text, the order of these resulted in four conditions (2 text languages x 2 orders of tests) which were counterbalanced across participants. To avoid confusion, all participants first received the text about the Sun, and then the text about sea otters. Combined with the factor interval, the experiment consisted of 12 conditions (2 x 2 x 3 factorial design).

In the first lab session, students studied four texts, starting with the texts about the Sun and sea otters (this study). Oral instructions were given in Dutch. At the start of the session, the students were informed that they had to study a text within a limited time frame of seven minutes and that they would be tested afterwards. They were not informed about a delayed recall test in the second session (and at the end of the second
session, we checked whether anyone studied the materials during the interval time, which was not the case). They were allowed to highlight sections of the texts or to make some sort of schematic summary, but only on the text itself, which they had to put aside once their study time was up. Testing time was ample with a 4-minute time limit to complete one test, to avoid individual differences in answering time. After the test phase, the procedure (study phase – test phase) was repeated for the second text. Next, they studied two short, expository handbook texts about experimental psychology in the same way, followed by a test of three open questions for every text respectively. At the end of the session, they were instructed to fill in all necessary tests (proficiency measures) online at home or during the lab time that was left in either session in order to receive their monetary reward and credits.

In the second session, students studied a longer academic text within a time limit of 15 minutes and took a test including open and multiple choice questions. Next, they filled in the long-term recall tests about the Sun and sea otters and carried out the operation span task.

Results

Scoring

The true/false judgments were scored dichotomously (correct/incorrect, with “I don’t know” as incorrect) with a correction key. The lure questions were analysed separately.

All data are available at https://osf.io/j8hay/ (Open Science Framework).

Testing whether the students were matched in the interval conditions

Because this study tests the effect of interval between-subjects, we first checked whether groups were matched on the control variables we assessed. Table 1 and 2 show that this was the case. There were no significant differences between the three groups if a Dunn-Šidák correction for multiple testing was taken into account (α = .002846). The mean L2 proficiency score of 74 for the English LexTALE is comparable to the previous
study (M = 72, Vander Beken & Brysbaert, in press) and is a typical score for this group of Dutch-English participants.

Participants in general had a higher reading motivation in L1 (M = 5.07, SD = 0.81) than in L2 (M = 4.46, SD = 0.89; Wilcoxon signed rank test resulted in V = 8362, p < .001), based on a sum score of several questions into reading attitude and motivation. Similar results from a previous study (Vander Beken & Brysbaert, in press) based on a single question are now confirmed with a more elaborate sum score in this study. The reliability of the objective measures was measured using Cronbach’s alpha, which was generally high. Table 3 displays the reliability measures and the correlations between the various measures. There were no motivational measures with M < 4, indicating that our participants were sufficiently motivated to take part in the experiment. When we asked their opinions about the usefulness of EMI at university, we found mildly positive scores as well (see Table 2).

Table 1. Mean scores of the interval groups on the various proficiency and intelligence tests (standard deviations between brackets).

<table>
<thead>
<tr>
<th>Tests</th>
<th>day group</th>
<th>week group</th>
<th>month group</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N = 49)</td>
<td>(N = 55)</td>
<td>(N = 51)</td>
<td>(N = 155)</td>
</tr>
<tr>
<td>Gender</td>
<td>36F/11M</td>
<td>43F/12M</td>
<td>39F/10M</td>
<td>118F/33M</td>
</tr>
<tr>
<td>Age</td>
<td>18.86 (3.49)</td>
<td>19.63 (5.16)</td>
<td>19.92 (4.43)</td>
<td>19.47 (4.44)</td>
</tr>
<tr>
<td>Dutch vocabulary (max = 75)</td>
<td>45.80 (7.59)</td>
<td>48.06 (8.53)</td>
<td>46.96 (7.62)</td>
<td>46.99 (7.94)</td>
</tr>
<tr>
<td>English LexTALE (max = 100)</td>
<td>72.63 (10.42)</td>
<td>74.95 (11.77)</td>
<td>73.83 (9.24)</td>
<td>73.85 (10.54)</td>
</tr>
<tr>
<td>English vocabulary size (max = 140)</td>
<td>95.71 (12.23)</td>
<td>95.75 (14.64)</td>
<td>96 (12.34)</td>
<td>96.09 (13.15)</td>
</tr>
<tr>
<td>Operation Span (WM) (max = 75)</td>
<td>58.02 (8.86)</td>
<td>57.18 (13.73)</td>
<td>59.74 (10.50)</td>
<td>58.29 (11.36)</td>
</tr>
</tbody>
</table>

Note: There were missing data points for 8 participants (re-running these comparisons with listwise deletion made no difference). There were no significant differences in the between-groups anovas for the continuous variables. The test statistics can be found at [https://osf.io/j8hav/](https://osf.io/j8hav/).
Table 2. Mean scores of the language groups on the self-ratings included in the questionnaire (standard deviations between brackets).

<table>
<thead>
<tr>
<th>Self-ratings</th>
<th>Day group (N = 49)</th>
<th>Week group (N = 55)</th>
<th>Month group (N = 51)</th>
<th>All (N = 155)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General motivation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test importance (7)</td>
<td>5.04 (1.24)</td>
<td>4.71 (1.36)</td>
<td>4.69 (1.319)</td>
<td>4.81 (1.27)</td>
</tr>
<tr>
<td>Performance vs. peers (7)</td>
<td>4.02 (0.80)</td>
<td>3.82 (0.86)</td>
<td>3.88 (0.55)</td>
<td>3.90 (0.75)</td>
</tr>
<tr>
<td><strong>Dutch academic reading</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitude (7)*</td>
<td>4.84 (1.17)</td>
<td>4.68 (1.23)</td>
<td>4.51 (0.90)</td>
<td>4.67 (1.11)</td>
</tr>
<tr>
<td>Intrinsic motivation (7)*</td>
<td>5.03 (0.98)</td>
<td>4.87 (0.95)</td>
<td>4.51 (0.90)</td>
<td>4.89 (0.97)</td>
</tr>
<tr>
<td>Total motivation (7)*</td>
<td>5.27 (0.77)</td>
<td>4.98 (0.78)</td>
<td>4.96 (0.85)</td>
<td>5.07 (0.81)</td>
</tr>
<tr>
<td><strong>English academic reading</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitude (7)*</td>
<td>5.77 (0.97)</td>
<td>5.65 (0.97)</td>
<td>5.53 (0.91)</td>
<td>5.65 (0.95)</td>
</tr>
<tr>
<td>Intrinsic motivation (7)*</td>
<td>4.50 (1.14)</td>
<td>4.35 (1.12)</td>
<td>4.20 (0.95)</td>
<td>4.34 (1.07)</td>
</tr>
<tr>
<td>Total motivation (7)*</td>
<td>4.67 (0.92)</td>
<td>4.4 (0.91)</td>
<td>4.30 (0.81)</td>
<td>4.46 (0.98)</td>
</tr>
<tr>
<td>Opinion about use of EMI (7)*</td>
<td>5.68 (0.95)</td>
<td>5.19 (1.30)</td>
<td>5.25 (1.15)</td>
<td>5.36 (1.16)</td>
</tr>
<tr>
<td><strong>Dutch language skill</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading (10)</td>
<td>9.02 (1.16)</td>
<td>9.2 (0.91)</td>
<td>8.94 (0.84)</td>
<td>90.6 (0.98)</td>
</tr>
<tr>
<td>Proficiency (10)*</td>
<td>9.07 (0.87)</td>
<td>9.02 (0.80)</td>
<td>8.84 (0.76)</td>
<td>8.98 (0.81)</td>
</tr>
<tr>
<td><strong>English language skill</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading (10)</td>
<td>7.67 (1.21)</td>
<td>7.67 (1.16)</td>
<td>7.46 (1.25)</td>
<td>7.60 (1.20)</td>
</tr>
<tr>
<td>Proficiency (10)*</td>
<td>7.49 (0.95)</td>
<td>7.37 (1.00)</td>
<td>7.16 (1.07)</td>
<td>7.34 (1.01)</td>
</tr>
</tbody>
</table>

Note: There were no significant differences in the Kruskall-Wallis tests to compare groups. Asterisks indicate sum scores. Likert-scale is indicated between brackets. The test statistics can be found at [https://osf.io/j8haw/](https://osf.io/j8haw/).
Table 3. Reliability and correlations of the proficiency and WM measures.

<table>
<thead>
<tr>
<th>Tests</th>
<th>Dutch voc. MC</th>
<th>Eng. LexTALE</th>
<th>Vocabulary size</th>
<th>Operation span</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch voc. MC</td>
<td>0.84</td>
<td>0.44</td>
<td>0.55</td>
<td>0.1</td>
</tr>
<tr>
<td>Eng. LexTALE</td>
<td>0.54</td>
<td>0.77</td>
<td>0.57</td>
<td>-0.01</td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>0.64</td>
<td>0.69</td>
<td>0.89</td>
<td>-0.06</td>
</tr>
<tr>
<td>Operation span</td>
<td>0.12</td>
<td>-0.01</td>
<td>-0.07</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Note: On the diagonal (in italic) is the cronbach’s alpha of each test. All numbers above that are original Pearson correlations. The numbers below the diagonal are the correlations corrected for reliability ($r_{xy}/\sqrt{r_{xx}r_{yy}}$). One participant did not fill in these four tests (N = 154). There were missing data points for 8 participants, which were omitted by pairwise deletion.

Performance on the memory tests

To analyse memory performance, the data set was analysed by means of mixed-effects logistic regression models with the lme4 package (version 1.1-7, Bates et al., 2014) of R (3.2.2) (R Core Team, 2015). Correctness of the answers was the binary output variable. Language (Dutch vs. English), interval (day/week/month), and session (immediate vs delayed) were included as categorical fixed effects. In a first model, we included the interactions between the three factors (language, interval, and session) and random intercepts and slopes for questions and participants. The R command we used was:

```r
glmer(correct ~ language * interval * session2 + (language+interval+session2|question) + (language+interval+session2|id), mydata, family="binomial", verbose=TRUE)
```

Table 5 displays the output of the analysis. The analysis indicated a significant effect of session ($\beta = -0.50$, SE = 0.16, $Z = -3.19$, $p < .01$), which means there was a lower chance participants remembered statements correctly in the second session. Furthermore, there were significant interaction effects between session 2 and the week interval ($\beta = -0.40$, SE = 0.20, $Z = -2.03$, $p < .05$; session 1 and day interval are the reference levels) and between session 2 and the month interval ($\beta = -0.63$, SE = 0.20, $Z =$
-3.10, p < .01), indicating that there was more forgetting the longer the interval between tests was. Importantly, language did not have a significant main effect and was not involved in significant interactions. Figure 1 illustrates the rate of forgetting based on the aggregated means in both languages (see Table 4 for the group means).

To check whether the full model hid a small language effect, we ran an extra model excluding the random slopes. This resulted in the same pattern of results. ANOVA comparison showed no significant difference in fit between the models. We also ran two models with only a main effect of language (i.e., without the interaction terms involving language) with and without random slopes. In these models there was no effect of language, making us confident that the absence of a significant effect of language is not due to us using a suboptimal model.

**Figure 1.** Observed average accuracy over items and participants in L1 (Dutch) and L2 (English). Note that the average on day 1 is aggregated over all interval groups (every participant was tested on day 1) while the other observations are based on separate groups. The SE was multiplied by 1.96 to obtain the confidence interval. More information on the procedure can be found in Brysbaert (2011, pp. 345-355).
Table 4. Percentage correct based on the aggregated scores per question.

<table>
<thead>
<tr>
<th></th>
<th>Immediate (all groups)</th>
<th>Day</th>
<th>Week</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>72.53</td>
<td>65.30</td>
<td>59.40</td>
<td>50.69</td>
</tr>
<tr>
<td>L2</td>
<td>71.47</td>
<td>64.68</td>
<td>55.00</td>
<td>46.95</td>
</tr>
</tbody>
</table>

Table 5. Output of the best fitted glmer-model of the memory scores.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Z value</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.21560</td>
<td>0.26067</td>
<td>4.663</td>
<td>3.11e-06 ***</td>
</tr>
<tr>
<td>LanguageEnglish</td>
<td>-0.14871</td>
<td>0.18023</td>
<td>-0.825</td>
<td>0.40932</td>
</tr>
<tr>
<td>Intervalmonth</td>
<td>-0.12344</td>
<td>0.18079</td>
<td>-0.683</td>
<td>0.49475</td>
</tr>
<tr>
<td>Intervalweek</td>
<td>0.06555</td>
<td>0.17244</td>
<td>0.380</td>
<td>0.70386</td>
</tr>
<tr>
<td>Session2</td>
<td>-0.49929</td>
<td>0.15630</td>
<td>-3.194</td>
<td>0.00140 ***</td>
</tr>
<tr>
<td>LanguageEnglish:Intervalmonth</td>
<td>0.09214</td>
<td>0.20549</td>
<td>0.448</td>
<td>0.65388</td>
</tr>
<tr>
<td>LanguageEnglish:Intervalweek</td>
<td>-0.05226</td>
<td>0.20132</td>
<td>-0.260</td>
<td>0.79518</td>
</tr>
<tr>
<td>LanguageEnglish:Session2</td>
<td>0.13311</td>
<td>0.19942</td>
<td>0.667</td>
<td>0.50447</td>
</tr>
<tr>
<td>Intervalmonth:Session2</td>
<td>-0.62578</td>
<td>0.20217</td>
<td>-3.095</td>
<td>0.00197 **</td>
</tr>
<tr>
<td>Intervalweek:Session2</td>
<td>-0.40191</td>
<td>0.19842</td>
<td>-2.026</td>
<td>0.04281 *</td>
</tr>
<tr>
<td>LanguageEnglish:Intervalmonth:Session2</td>
<td>-0.21029</td>
<td>0.27476</td>
<td>-0.765</td>
<td>0.44404</td>
</tr>
<tr>
<td>LanguageEnglish:Intervalweek:Session2</td>
<td>-0.11609</td>
<td>0.27599</td>
<td>-0.421</td>
<td>0.67402</td>
</tr>
</tbody>
</table>
Memory illusions

The questions that were added to induce memory illusions were repeated in both sessions and for that reason they were excluded from the general memory performance analysis. If the first test induced memory illusions, more incorrect answers are expected on these questions in the second test. Table 6 displays the percentage of false alarms (“yes”-answers) in all conditions based on 5 or 6 questions per test. These averages clearly show that more false memories arise after a longer interval, despite the option to answer “I don’t know”. To analyse performance, the data subset was analysed by means of mixed-effects logistic regression models. The binary output variable corresponded to the presence of a false alarm. Language, interval, and session were included as categorical fixed effects. Fitting the full model with all interactions between the three factors (language, interval, and session) and random intercepts and slopes for questions and participants failed due to a convergence error. This is probably due to a lack of variance since most answers were “no”-answers or “I don’t know”-answers, both zero values. The data could only be analysed using a glmer-model with the interaction between session and interval, without any other interactions and random slopes. In this model, there was a significant main effect of language ($\beta = 0.44$, SE = 0.11, $Z = 4.16$, $p < .001$) and session ($\beta = 0.40$, SE = 0.19, $Z = 2.01$, $p < .05$), so there was a higher chance of false alarms in English compared to Dutch and in the second session compared to the first. There were significant interactions between session and the week interval ($\beta = 1.05$, SE = 0.26, $Z = 3.97$, $p < .001$) and session and the month interval ($\beta = 1.48$, SE = 0.26, $Z = 5.73$, $p < .001$), which means the probability of false alarms increases after longer intervals Table 7 displays the output of the analysis.

Table 6. Percentage of false alarms (illusions) based on the aggregated scores per lure question (“yes”-answers). Note: this is based on 5 or 6 questions per test.

<table>
<thead>
<tr>
<th></th>
<th>Immediate (all groups)</th>
<th>Day</th>
<th>Week</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>8.66</td>
<td>15.77</td>
<td>24.04</td>
<td>40.32</td>
</tr>
<tr>
<td>L2</td>
<td>14.52</td>
<td>15.32</td>
<td>34.21</td>
<td>42.62</td>
</tr>
</tbody>
</table>
Table 7. Output of the best fitted glmer-model of the memory illusion scores.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Z value</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.67848</td>
<td>0.34797</td>
<td>-7.698</td>
<td>1.39e-14 ***</td>
</tr>
<tr>
<td>LanguageEnglish</td>
<td>0.43880</td>
<td>0.10539</td>
<td>4.164</td>
<td>3.13*e-05 ***</td>
</tr>
<tr>
<td>Intervalmonth</td>
<td>0.11819</td>
<td>0.22569</td>
<td>0.524</td>
<td>0.6005</td>
</tr>
<tr>
<td>Intervalweek</td>
<td>-0.08874</td>
<td>0.22472</td>
<td>-0.395</td>
<td>0.6929</td>
</tr>
<tr>
<td>Session2</td>
<td>0.38980</td>
<td>0.19371</td>
<td>2.013</td>
<td>0.0442       *</td>
</tr>
<tr>
<td>Intervalmonth:session2</td>
<td>1.48220</td>
<td>0.25860</td>
<td>5.372</td>
<td>9.94e-09 ***</td>
</tr>
<tr>
<td>Intervalweek:session2</td>
<td>1.04588</td>
<td>0.26333</td>
<td>3.972</td>
<td>7.14e-05 ***</td>
</tr>
</tbody>
</table>

Discussion

In this experiment, we tested students’ recognition memory in Dutch (L1) and English (L2) on an immediate and delayed test, using true/false judgment items from a previous study (Vander Beken & Brysbaert, in press). Since participants were divided in groups to determine the time of the delayed test, those groups were compared on several objective and subjective measures of proficiency and motivation. There was no difference between groups.

As expected, we did not find an L2 recognition cost on the initial test. Since languages were directly compared in a within-participant design, this robustly confirms the results of the previous study. On the delayed test, there was no significant language effect either. Two conclusions can be drawn from this observation. Firstly, for education, this means that it is no disadvantage for students to be tested on the long term in English, at least for recognition memory. There seems to be no loss of information even though study time was the same in both languages. Secondly, we have found no indication of a disadvantage situated at the level of storage of the mental model and, thus, no evidence for the weaker-links hypothesis.

There are several possible explanations for this finding. It is possible that the recall deficit from the previous study is located at the production level only, which means that people do not remember less in L2 but have more difficulty writing up their recalled memories in L2. Of course, not being able to express the knowledge you have can also be
No L2-disadvantage in long term recognition memory

problematic. To confirm this possible explanation, a cross-lingual study (with L2 text – L1 test but also L1 text – L2 test conditions) should show a clear disadvantage of the translation from concept to L2 wordings in all L2 production conditions.

On the other hand, one could argue that a weaker mental model with less “rich” memory traces could still account for unaffected recognition and that memory traces that are weaker produce marginal knowledge. Still, the levels-of-processing framework does suggest that “elaboration coding”, i.e. deeper processing with more semantic analysis, results not only in a more elaborate but also a longer lasting memory trace (Craik & Lockhart, 1972). If it were indeed the case and our participants encoded more surface information and less semantic information in the non-dominant language, we would probably have found some long-term memory loss in L2.

These two views seem to exclude one another. Nevertheless, the results from this paper can actually be explained by a combination of opposite effects as well. If the encoding specificity principle can have an effect on studying or reading texts, the unusual context of an L2 study text might create strong contextual cues for retrieving information, compared to L1. In addition, students process less information in L2 than in L1 in the time between the immediate and delayed test, yielding a larger uniqueness of the memory trace in L2 than in L1. So if the encoding of information was less deep in L2 than L1 and the memory trace less strong as a consequence, then there would be a weaker trace in the first instance that suffers less from information interference and is more easily retrieved in a second stage. However, it would be a large coincidence if these two effects were of the exact same size, resulting in a null effect.

Interestingly, Francis and Gutiérrez (2012) showed that the levels-of-processing effect is smaller in L2 than in L1, meaning that shallow processing tasks (e.g. word recognition) yield better recognition performance in the weaker language, but that this advantage decreases for deeper encoding tasks. In other words: deeper encoding tasks mainly improve L1 performance. Taking into account that understanding and remembering a text is a more demanding task in general, this pattern of results is very similar to the combined results from this study and the previous one. The authors explain their observations by a combination of weaker links and resource limitations (Francis & Gutiérrez, 2012).
The experiment also included an attempt to induce memory illusions. More illusions on the delayed test in L2 would suggest that the memory trace is indeed weaker due to shallower processing. We found a main effect of language, indicating that more false memories arise in English. The effects of interval and session, and their interaction, simply suggest forgetting over time. Due to low variance, it was impossible to investigate whether the language effect is mediated over time and whether the memory trace actually fades out more easily. Still, the finding that more lures are remembered as correct in English might be explained by the levels-of-processing effect. If processing is shallower in L2, then maybe a false statement interferes as a new, unique memory instead of being rejected based on the contents of the text. This would indicate weaker encoding rather than storage and indirectly strengthens our hypothesis. Nonetheless, despite lower performance on these questions compared to the other questions, we cannot be one hundred percent sure that our attempt to yield false memories with this construction was successful (usually this is tested with multiple choice questions of which one answer is a lure) and we should be careful with strong conclusions about this exploratory element of this study.

To conclude, in this experiment, we found no clear-cut evidence for the weaker-links hypothesis. If L2 memory showed a higher forgetting rate, we would conclude that semantic links are weaker, resulting in weaker memory traces, but this is not the case. Following the logic of Francis and Gutiérrez (2012), the results from this and the previous study could possibly be explained within the levels-of-processing framework: shallow processing tasks on word level result in better L2 recognition than L1 recognition, but there is no such L2 advantage for deeper processing tasks. If you take into account that studying a full text is a more complex task and requires more resources during encoding (according to the resource hypothesis), this effect could translate to our findings. The levels-of-processing effect is larger in a non-dominant language, so perhaps at text level, shallow processing of the L2 texts results in unimpaired long-term recognition performance, but the necessary deeper processing in L2 fails to some extent, compared to L1. (Note that this could still arise within a weaker links framework). Furthermore, we did not find any recognition cost in L2, suggesting that students can be tested in L2 with recognition tests without risking an underestimation of their (possibly marginal) knowledge. In other words, as far as recognition memory goes, the cost-
effectiveness of education is not endangered by EMI: the acquired knowledge is retained over a long retention interval (see the introduction section and Berger et al., 1999, p. 438). Further research is necessary to confirm whether the L2 recall cost is actually a production deficit or whether the reason for this disadvantage is more complex and located at the encoding stage of memory. It would also be of great value to explore the current research line with various tests, intervals, and types of bilingual information retention, to discover the commonalities and contrasts between L1 and L2 memory.
References


STUDYING TEXTS IN A NON-NATIVE LANGUAGE:
A FURTHER INVESTIGATION
OF FACTORS INVOLVED IN THE L2 COSTS

Heleen Vander Beken, Ellen De Bruyne,
& Marc Brysbaert

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 (Vander Beken & Brysbaert, in press), not even on delayed tests after a month (Vander Beken, Woumans, & Brysbaert, in press). 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 conducted three experiments in L1-L1, L2-L1 and L2-L2 conditions with cued recall (open/closed questions) and multiple choice questions. On the one hand, no significant language differences were found for cued recall of short texts in Experiment 1. On the other hand, a significant disadvantage was found for studying a text in L2 compared to L1 on cued recall and multiple choice questions in L1 (Experiment 2). These results indicate that both the encoding and reproduction processes are sometimes impaired in L2 and that the effect (size) depends on several factors such as time, task requirements, and proficiency level.
Introduction

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.

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 students at Ghent University. In the first study (Vander Beken & Brysbaert, in press) we focused on recognition memory versus recall for short expository texts. We observed a large RECALL COST 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 for the testing phase), so that some of the recall cost could be due to a shortage of writing time. To sum up: A large difference was found due to the use of a non-native language in a recall protocol, but no difference was found in a recognition test.

In the second study (Vander Beken, Woumans, & Brysbaert, in press), we tested recognition memory at several time intervals: immediately after studying the text, and either one day, one week, or one month later. No additional forgetting was observed in L2 compared to L1, even not after a month. This way, we examined one possible origin of the recall cost: the strength of memory trace in L2 versus L1. Free recall involves an
active process of retrieval for which no external cues are available, based on the task only. In a yes/no recognition task, cues are given. Some knowledge that has not been stored so actively or successfully that it is available for immediate and deliberate access, can be triggered and found with the help of cues. As a consequence, it will not emerge in a free recall protocol. In memory research, this type of knowledge is called marginal knowledge (Berger, Hall, & Bahrick, 1999; Cantor, Eslick, Marsh, Bjork, & Bjork, 2014). Perhaps studying in L2 leads to an unstable memory trace, in which the individual elements can be retrieved when cued, but not recalled without help. This disparity could be caused by the fact that information is encoded more superficially in L2. Such an explanation can be derived from 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 a second study (Vander Beken, Woumans & Brysbaert, in press) was that stronger memory traces, the result of deeper processing, also predict longer retention. In other words: If the recall cost in L2 is due to impaired encoding, we expect a steeper forgetting curve in L2 than in L1, because the knowledge decays at a higher rate. This must be true especially for the proportion of knowledge which is recognised but not recalled (the marginal knowledge). Therefore, the best choice to investigate this is the yes/no recognition task. In addition, the evidence should be particularly clear for that task because performance in both languages is the same immediately after studying the text. As was mentioned earlier, forgetting in L1 and L2 was similar, contrary to our predictions.

In the present series of studies, we address a second possible interpretation of the recall cost, namely that it relates to the requirement of text production. In free recall, participants not only have to retrieve the information but also reproduce 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 also be the origin of the discrepancy between recall and recognition. To investigate this possibility, the obvious next step is to ask participants to study a text in L2 and recall it in L1 or in L2: if a production issue is the origin, L1 recall of an L2 text will result in equal performance as L1 recall of an L1 text.
However, the outcome of such a test also depends on the assumption one makes about how L1 and L2 texts are stored in memory. The dominant and rather intuitive view is that meaning is stored at a separate, abstract level free of language, (Alba & Hasher, 1983; Dijkstra & van Heuven, 2002; Schank, 1972, 1980). One of the arguments for this view is that meaning in bilinguals must be shared between the languages. According to this view, studying in L1 and L2 results in information stored in the same language-independent code, and any language difference is due to encoding or retrieval. On the other hand, it cannot be excluded that text memory is (at least partly) language-dependent. In such a situation there may be an advantage if both encoding and retrieval occur in the same language. This idea is based on the encoding-specificity principle (Tulving & Thomson, 1973). Experimental evidence of this principle in the language domain has indeed 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).

In the literature on L2 text memory, the issue of higher L2 load vs. encoding specificity is confounded in nearly all experimental findings. In some studies, the test language at retrieval is congruent with the text language at encoding (e.g. Connor, 1984; Donin, Graves, & Goyette, 2004; Vander Beken & Brysbaert, in press), in others, the memory test is always presented in L1 (fixed) (e.g. Brantmeier, 2005; Chang, 2011; Davis, Lange, & Samuels, 1988; Joh, 2006; Roussel, Joulia, Tricot, & Sweller, 2017), resulting in a cross-lingual situation. Retrieval in L1 avoids an effect due to inferior L2 production skills, which is a widely spread argument for this methodological choice (e.g. Roussel et al., 2017), but it means that there is a language switch between encoding and retrieval, which may affect the score by hampering memory retrieval.

Very few papers systematically compared what happens in L2-L2, L2-L1 and L1-L1 conditions with the same materials, which is needed if we want to get a grasp of the nature of the recall cost (an L1-L2 condition is not included because it is ecologically invalid and it would be impossible to know whether low scores are due to insufficient L2 vocabulary knowledge or a different matter). In two studies (Gablasova, 2014, 2015), the issue was addressed (or controlled for) by collecting half of the answers in L1 and half in L2. However, the author did not systematically look into the differences. In another study, Horiba and Fukaya (2015) looked into the effects of topic familiarity and
proficiency on discourse recall both in L2-L2 and L2-L1, reporting better content recall in L1. However, they did not include a unilingual L1-L1 condition.

A study that did adopt the approach with all three language conditions was conducted by 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 L2. For the other text read in L2, they were asked to recall in L1. Surprisingly, the quality of recall did not differ between the language conditions, although there was a trend showing lower L2 recall than L1 recall of the text studied in L2. However, the study was underpowered, with a total of 38 participants (divided in groups based on study background and proficiency level with language as a within-subject factor), which may explain the non-significance of the trend. In addition, since recall was oral, the L2 production cost may have been less than in written recall (in which spelling, for example, is an additional requisite for production).

Taking everything together we hypothesize that, if the L2 production cost account is correct, we should find a large recall disadvantage in L2-L2 (replicating Vander Beken & Brysbaert, in press) but a smaller or absent disadvantage in L2-L1 relative to L1-L1.

To be able to conclude anything with reasonable certainty, it is important to make sure that the study is properly powered. Because such studies require access to many participants (and are also time-consuming), this design was combined with the Vander Beken et al. (in press) study, testing students from the same participant pool, and repeated measures designs were used were possible. The present experiments were run after the participants finished the immediate test or in the second session of Vander Beken et al. (in press), in order not to influence the results of that study. The fact that the same participants were tested means we were forced to use different and unrelated texts as in Vander Beken & Brysbaert (in press) and Vander Beken et al. (in press). This has both advantages and disadvantages. The main drawback is that the data may deviate from the previous studies because other texts were used. The main advantage is that we can look at the generalisability of the results, dealing with the risk of drawing strong conclusions based on the same two short, expository texts (one on sea otters and one on
the Sun). The new text and test materials were more diverse than the ones used in the previous studies. In addition, we used a new range of tests to home in on the recall versus recognition distinction.

The current study consists of two experiments. The Experiment 1 investigated the three important language conditions (L1-L1, L2-L1, L2-L2) using short expository texts and open questions, tapping into recall in a lab situation very similar to real-life university examination. Experiment 2 focused on a longer scientific text (a part of an academic article) and tested both recognition memory via a multiple choice test and recall via open questions. Due to time and feasibility constraints, the latter experiment only used one text and was thus limited to the language conditions L1-L1 vs L2-L1.

Experiment 1: Cued recall of handbook paragraphs in L1-L1, L2-L1, and L2-L2

Method

Participants

An effect size of $d = .4$ or more has practical implications in applied settings such as education (Ferguson, 2009). Assuming a standardized effect size of $d = .4$, this entails that a group of 52 participants must be tested in a repeated measures design, at least when each participants reads only one text per condition (Brysbaert & Stevens, in press).

The data in this chapter were collected in a subgroup of the participants tested in Vander Beken et al. (in press). All participants were first year psychology or educational sciences students at Ghent University, participating in partial fulfilment of course requirements combined with a fixed financial reward. Almost all had Dutch as their dominant language (and were otherwise excluded from analysis). Note that in this study L1 was defined in terms of dominant language, not as the first acquired language. 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. In some of the university courses English handbooks are used as well, even though the teaching happens in Dutch. Students with reading problems were excluded, as well as students who did not complete all tests or did not follow the instructions correctly (e.g., recalled in another language than instructed).

All in all, data of 53 participants were retained in experiment A (from an initial number of 68) and 63 in experiment B (from a total of 78).

Text materials

In previous experiments, we conscientiously tried to match texts between languages on content and word frequency measures. A downside of this approach is the risk that the translations are experienced as less natural than the original, despite the fact that they were checked and double-checked. Therefore, we decided to step away from strict, potentially unrealistic experimental control and to opt for materials as close to real-life situations as we could find. We went looking for texts that existed both in Dutch and English. In addition, we wanted the texts to be real educational material for students, to increase the ecological validity of the experiment. From two psychology books, written by the same author, we selected two excerpts that were very similar (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 we selected were not studied in the students’ courses yet, and we tested them early in the academic year to avoid strong prior knowledge.

Two texts were extracted from the books. The first was called 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. See Appendix B at the end of this dissertation for the texts.

**Retrieval 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 questions were answered independently by two of the authors of the present chapter, one of which is the author of the books, to compare the answers, with the purpose of creating a robust correction key (and in addition, checked by the second rater of the retrieval tests). 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 Appendix B.

**Scoring**

As indicated above, a correction key was developed for all questions, dividing the answers into elements that were scored separately. All memory tests were scored by the first author and a second rater affiliated to the Department of Experimental Psychology (thus having sufficient knowledge of the topic), using the key. 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.
Control variables

We collected proficiency and working memory data for all participants. They took two English vocabulary tests, the lexical decision test LexTALE (Lemhöfer & Broersma, 2012) and the multiple choice Vocabulary Size Test (Nation & Beglar, 2007; 14,000 version), both aimed at non-native speakers of English. The L1 receptive vocabulary knowledge was estimated with a 75-item Dutch multiple choice vocabulary test developed at the department (see Appendix C). Furthermore, we administered the automated Operation Span task (Unsworth, Heitz, Schrock, & Engle, 2005) to measure working memory capacity. Finally, a reading motivation questionnaire was administered to examine general motivation, language-specific motivation and attitudes (see Appendix D). Note that the control measures are also reported in Vander Beken et al. (in press). Only the group upon which they are calculated is smaller here.

Table 1 shows the means and standard deviations on the various tests. As can be seen, L2 proficiency was substantially lower than L1 proficiency. The level of L2 proficiency, as measured with LexTALE (M = 74), is comparable to that in previous studies with similar participant groups (see Vander Beken & Brysbaert, in press, for more information). 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, 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 1A and 1B, simplifying the comparison of the experiments.
Table 1. Mean scores on the various proficiency and intelligence tests of the participants in Experiments 1A and 1B (standard deviations between brackets).

<table>
<thead>
<tr>
<th>Tests</th>
<th>1A</th>
<th>1B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>39F/13M (1 NA)</td>
<td>47F/14M (2 NA)</td>
</tr>
<tr>
<td>Age</td>
<td>19.05 (1.75)</td>
<td>20.30 (5.06)</td>
</tr>
<tr>
<td>Dutch vocabulary MC (max = 75)</td>
<td>46.55 (7.55)</td>
<td>47.92 (7.60)</td>
</tr>
<tr>
<td>English LexTALE (max = 100)</td>
<td>73.54 (9.39)</td>
<td>73.95 (10.80)</td>
</tr>
<tr>
<td>English Vocabulary Size (max = 140)</td>
<td>95.68 (11.83)</td>
<td>95.94 (13.38)</td>
</tr>
<tr>
<td>Operation Span (WM) (max = 75)</td>
<td>57.30 (13.17)</td>
<td>58.48 (11.07)</td>
</tr>
<tr>
<td>Self-rating of Dutch lang. skill (max = 10)</td>
<td>9.6 (0.88)</td>
<td>8.89 (0.75)</td>
</tr>
<tr>
<td>Self-rating of English lang. skill (max = 10)</td>
<td>7.57 (1.07)</td>
<td>7.11 (1)</td>
</tr>
</tbody>
</table>

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

**Design**

Since we had only two similar excerpts from the two different handbooks at our disposal, we split the study in two smaller experiments with one repeated-measure variable. In Experiment 1A, 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 1B, 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. Participants were not informed beforehand about the language of the test they would get, but they were informed they had to answer in the same language as the test, even when that language was different from the text. Notice that the L2-L1 condition was the same in both experiments (and hence can be compared across experiments).
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., in press). Next, 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, an Open Source web application available through the university. There was only a short interval 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.

In the second session, students studied a longer academic text within a time limit of 15 minutes and took a test including open and multiple choice questions, which is reported in Experiment 2. Next, they filled in the delayed recall tests about the biology topics (Vander Beken et al., in press) and completed the operation span task, programmed in E-Prime 2.0.10 (Unsworth et al., 2005).
Table 2. Means and standard deviations (between brackets) of the memory scores in experiment 1 (correct points out of 5).

<table>
<thead>
<tr>
<th>Experiment 1A: all tests in L1</th>
<th>Experiment 1B: all texts in L2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L1 text</strong></td>
<td><strong>L1 test</strong></td>
</tr>
<tr>
<td>3.07 (1.15)</td>
<td>2.82 (1.20)</td>
</tr>
<tr>
<td><strong>L2 text</strong></td>
<td><strong>L2 test</strong></td>
</tr>
<tr>
<td>2.77 (1.15)</td>
<td>2.63 (1.24)</td>
</tr>
</tbody>
</table>

Results

Table 2 displays the mean memory scores and their standard deviations in Experiment 1A and 1B. Data files are available at [https://osf.io/c67ya](https://osf.io/c67ya).

Memory scores in Experiment 1A

Based on the average scores, the mean was 3.07 (SD = 1.15) for the L1-L1 test and 2.77 (SD = 1.15) for the L2-L1 test (see Table 2 and Figure 1). The data were right-skewed, but since our sample size was sufficiently large (N = 53), a paired samples t-test was used for analysis. There was no evidence that the scores were higher in the L1-L1 conditions than in the L2-L1 condition (t(52) = 1.588, p = 0.12). A post-hoc calculation of effect size resulted in a Cohen’s d estimate of d = 0.20, which is considered a “small” effect. The confidence interval of the effect size ranged from -0.19 to 0.59. Next to hypothesis testing, from which we cannot refute the null hypothesis, we also ran a Bayesian analysis to assess to which extent the observed distribution corresponds to expected distribution of the null hypothesis. The Bayes factor B_{10} was 0.49, which denotes anecdotal evidence for the null hypothesis.

\[ \text{To make sure, we also ran a non-parametric test. The Wilcoxon signed-rank test did not yield different results from the T-test.} \]
**Factors involved in the L2 costs**

Memory scores in Experiment 1B

Mean recall was 2.82 (SD = 1.2) for L2-L1 and 2.63 (SD = 1.24) for L2-L2 (see Figure 2). 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 B₁₀ was 0.192 here, which denotes moderate evidence for the null hypothesis.

![Figure 1. Memory scores per language condition on 5 points in experiment 1A (all tests in L1), with 95% confidence intervals.](image-url)
Figure 2. Memory scores per language condition on 5 points in Experiment 1B (all texts in L2) with 95% confidence intervals.

Discussion

In the first experiment, 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. No significant effect of language was found. When using a power of 80% to detect an effect of $d = .4$ (a practically relevant effect size), no convincing evidence was found that cued recall of expository texts is easier in L1-L1 than in L2-L1, or in L2-L1 than in L2-L2. At the same time, there was twice a trend in the direction of L1-L1 > L2-L1 > L2-L2, similar to what was observed by Chen and Donin (1997). Based on post-hoc calculations of effect size, we cannot exclude the possibility that there are effects of such a small size that our study was not powered enough to detect it. If both effects are additive, this may indicate that there would be a small to intermediate effect in an L1-L1 versus L2-L2 design, which

---

Though prudence is needed when juxtaposing separate experimental groups, the L2-L1 condition which appeared in both experiments was very similar and provides a reference point for a comparison.
is similar to the findings in Vander Beken and Brysbaert (in press). The type of task might mediate the effect size, with open questions in this study yielding a smaller effect than the free recall task. Though the effect does not reach the threshold for effects that are practically relevant in education as proposed by Ferguson (2009), these smaller effects can still be informative for large-scale decisions, such as the choice for the use of a foreign language as the medium of instruction in higher education.

Another interesting observation is that we found no evidence for the encoding-specificity principle, according to which L2-L2 should have resulted in better performance than L2-L1. So, the L2 production cost seems to outweigh any encoding-specificity benefit that might be present. In general, it seems that every addition of L2 in the task causes a minor drop in performance. Arguably, this points towards some difficulties both at the encoding or storage level and at the production level. In the general discussion, we will go deeper into the theoretical implications of this finding. But first, Experiment 2 is reported, in which participants read an academic text either in L1 or in L2 and received a cued recall and multiple choice test in L1.

**Experiment 2: Cued recall and recognition of an academic article**

**Method**

**Participants**

Assuming a practically standardized effect size of $d = .4$ (Ferguson, 2009), this entails that two groups of 99 participants must be tested in a between-subjects design. So
ideally, 198 would be tested. In this experiment, data was collected for 124 participants\(^3\). This means we required an effect size of \(d = .5\) to reach 80% power. For an effect size of \(d = 4\), we only had 67% power. These participants were part of the same group that was tested in Experiment 1 (see above).

**Text materials**

In this experiment, we selected another type of text that is very relevant to higher education students: an academic article. The text was selected from an earlier study by De Bruyne and Valcke (in preparation) and was a 4-page fragment from *Cognitive, metacognitive, and motivational aspects of problem solving* (Mayer, 1998). The topic is studied in the educational sciences department at our faculty and was used to investigate L2 reading strategies in first-year educational sciences students. De Bruyne and Valcke (in preparation) translated the text to Dutch. The English version was 1749 words long; the Dutch version 1760. The text consisted of one introductory paragraph and three sections with a section title.

**Test materials**

Two types of tests were created for this test: a test with 5 open/closed questions and one with 14 multiple choice questions, partially taken from De Bruyne and Valcke (in preparation) and partially created for the purpose of the present study. The questions were not constructed for direct comparison; hence the different number of questions of each type. They were constructed to tap into different skills (reproducing or explaining versus recognition). All participants received both tests in a single test session. The tests were in Dutch; an English translation can be found in Appendix B.

**Control variables**

The same control variables as in Experiment 1 were used. Table 3 shows the results. Since the test language was a between-participants variable in the present experiment, it was even more important that the two groups were equivalent on the control variables. There were no significant differences between the two groups if a Dunn–Šidák correction

---

\(^3\) This was the consequence of an error in our initial sample size calculation which we unfortunately noticed after testing. Still, power is sufficient to detect an effect of \(d = .5\), worthwhile to report after detecting an L2 recall cost of \(d = .86\).
for multiple testing was used. Since group differences were tested on 7 variables, the alpha level is corrected to .0073008. Without the correction the difference in Dutch vocabulary scores would exceed the p = .05 threshold. We think the corrected p-level is the one to be used. Even without correcting for multiple testing, there are no group differences for L2 proficiency.

### Table 3. Mean scores on the various proficiency and intelligence tests of the participants in Experiment 2 (standard deviations between brackets).

<table>
<thead>
<tr>
<th>Tests</th>
<th>L1 group</th>
<th>L2 group</th>
<th>T-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>47F/8M (1 NA)</td>
<td>46F/20M (2 NA)</td>
<td>0.67897</td>
</tr>
<tr>
<td>Age</td>
<td>19.99 (6.05)</td>
<td>19.41 (2.58)</td>
<td>1.8353</td>
</tr>
<tr>
<td>Dutch vocabulary MC (max = 75)</td>
<td>48.86 (8.18)*</td>
<td>46.19 (7.87)*</td>
<td>1.4684</td>
</tr>
<tr>
<td>English LexTALE (max = 100)</td>
<td>75.15 (10.39)</td>
<td>72.44 (10.04)</td>
<td>-0.3061</td>
</tr>
<tr>
<td>English Vocabulary Size (max = 140)</td>
<td>95.05 (14.27)</td>
<td>95.80 (12.62)</td>
<td>-0.9131</td>
</tr>
<tr>
<td>Operation Span (WM) (max = 75)</td>
<td>56.68 (12.86)</td>
<td>58.63 (10.81)</td>
<td>-1738.5</td>
</tr>
<tr>
<td>Self-rated Dutch lang. skill (max = 10)</td>
<td>8.94 (0.96)</td>
<td>9.00 (0.86)</td>
<td>2023.5</td>
</tr>
<tr>
<td>Self-rated English lang. skill (max = 10)</td>
<td>7.41 (1.03)</td>
<td>7.27 (0.92)</td>
<td>0.8076</td>
</tr>
</tbody>
</table>

Note: Underlined variables indicate sum scores. The asterisk refers to a significant difference at an alpha-level of .05, but none of the group comparisons indicated a significant difference after a Dunn-Šidák correction for multiple tests was taken into account (α = .0073008). The test statistics can be found at [https://osf.io/c67ya](https://osf.io/c67ya).
Design

Because of the length of the text, all students only received one text, either in Dutch (L1) or in English (L2). They received both open questions and multiple choice questions in Dutch and were asked to answer in Dutch (L1). This resulted in a 2 x 2 mixed factorial design with test type as a repeated measure and language of text as a between-subjects factor.

Procedure

The sequence of tasks discussed in Experiment 1 also applies to the present experiment.

Oral instructions were given in Dutch. The participants were informed that they had to study a text within a 15-minute time limit, and would receive a Dutch test afterwards. They were allowed to highlight sections of the texts or to make some sort of schematic summary, but only on the text itself, which they had to put aside once their study time was up.

The text was presented on paper, in Times New Roman, font size 10, in two or three paragraphs per page (four pages in total). The test was administered online using LimeSurvey. A time limit of 10 minutes was used for the part of the test with open questions, to ensure that students continued to the next part and finished within lab time. Between text and test there was a short interval as in experiment 1.

Results

Scoring

First, a correction key was developed for each question. All open questions were scored by the first and the second author, using that key. Since we were interested in memory retrieval, rather than writing skill, we again adopted the guideline for PISA tests (see appendix in Cartwright, 2012) not to punish spelling and grammatical mistakes
unless they obscured meaning. Every part of the answers was scored dichotomously (so, the answer to the question “what are the three components of problem solving?” received a score of 0 or 1 for each of the three components). The mean interrater reliability over all (sub)items was $r = .76$. The scores were first averaged per item (so that, for the example question, the average of the three elements was taken) and then averaged across the questions per question type (open question and multiple choice).

**Memory performance**

Based on the proportions correct, the mean score for open questions was 0.32 (SD = 0.25) in the L1-L1 condition and 0.25 (SD = 0.20) in the L2-L1 condition. For the multiple choice questions, the mean proportions were 0.45 (SD = 0.15) in the L1-L1 condition and 0.37 (SD = 0.14) in the L2-L1 condition. See table 4 and Figure 3 for an overview of these results.

Table 4. *Means and standard deviations (between brackets) of the memory scores in experiment 2 (proportion).*

<table>
<thead>
<tr>
<th></th>
<th>Dutch (N = 56)</th>
<th>English (N = 86)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Open questions</strong></td>
<td>0.32 (0.25)</td>
<td>0.25 (0.20)</td>
</tr>
<tr>
<td><strong>Multiple choice</strong></td>
<td>0.45 (0.15)</td>
<td>0.37 (0.14)</td>
</tr>
</tbody>
</table>

Note that this scoring did not affect the language effect reported in the *Memory performance* section. Summing the scores to 1 point instead of 3 creates a total score that reflects the total understanding better: otherwise, these sub-questions would have an excessive weight on the total score. Nevertheless, we reran the analyses leaving all sub-components at a score of 1 instead of summing them, to verify that the language effect was no involuntary consequence of the weighted compound score.
To analyse differences in memory performance, we used a 2 (language group) × 2 (question type) mixed factorial type III ANOVA. Each participant had only been presented with one text to study (either in L1 or in L2) but received both open questions and multiple choice questions, so question type was a repeated measure. The analysis indicated a significant main effect of language ($F(1,122) = 8.643, p < .01$, $\eta^2_p = .07$), a significant main effect of question type ($F(1,122) = 36.615, p < .001$), but no significant interaction between both variables ($F(1,122) = 0.042, p = .84$). The size of the language

Note that the effect of question type is related to the way the tests were constructed: the open questions that required multiple elements in the response were averaged to a score of 1 (if these would not have been averaged, the compound score of the open questions would be higher, resulting in a smaller effect of question type).
effect corresponded to Cohen’s $d = .40$, with a confidence interval of 0.15–0.66. In correspondence to the previous experiment, we ran a Bayesian analysis as well. The Bayes factor ($BF_{10}$) for a model with language was 6.68, which is moderate evidence for a language effect (compared to no such effect). The Bayes factor for a model with question type was $1.14e^6$, which is extreme evidence for an effect of question type (compared to no such effect). Furthermore, the analysis confirmed that the model including a main effect for language and question type was the best model, and was better than a model including the interaction. See table 5 for the output of the analysis.

<table>
<thead>
<tr>
<th>Model</th>
<th>$B_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language + ID</td>
<td>6.68</td>
</tr>
<tr>
<td>Question Type + ID</td>
<td>$1.14e^6$</td>
</tr>
<tr>
<td>Language + Question Type + ID</td>
<td>$9.88e^6$</td>
</tr>
<tr>
<td>Language + Question Type + Language * Question Type + ID</td>
<td>$1.89e^6$</td>
</tr>
</tbody>
</table>

### Discussion

In the second experiment, we presented participants with a 1,750 word text taken from an academic review article. The text was presented in L1 or L2. Participants were asked to respond to open questions and multiple choice questions in L1. We observed a significant effect of the language of the text irrespective of the types of questions. A language effect for open questions replicates the effect reported by Vander Beken and Brysbaert (in press), although it is smaller in effect size ($d = .4$ instead of $d = .8$). Surprisingly, the language effect observed here remains for the multiple-choice questions, although this type of question does not involve language production and
arguably taps more into recognition than recall. Two differences may be responsible for
the observation. First, the multiple-choice questions used in the present experiment may
have tapped more into recall processes than the true/false questions used in the previous
experiments (Vander Beken & Brysbaert, in press; Vander Beken et al., in press). If this is
the case, we would predict an even larger cost for the multiple-choice questions in an L2-
L2 condition. A second factor that may be involved, however, is the time participants had
to study the texts. In the present study, they had only 15 mins for 1,750 words, whereas in
the previous studies they had 7 mins for 300-500 words. The overall scores reflect that
either time for reading was too short or the questions were too difficult, since the mean is
not even above 50% (though the mean score for multiple choice questions is still above
the 25% chance level). The shorter study time may have resulted in a more pronounced
L2 disadvantage (together with higher demands of the text, possibly).

Notice that the encoding-specificity principle may also have contributed to the
findings of Experiment 2, as the languages of text and test were the same in the L1-L1
condition but not in the L2-L1 condition.

General discussion

In this study, we continued our investigation into possible costs of having to study in
a non-native language (more specifically, in English). In previous studies (Vander Beken
& Brysbaert, in press; Vander Beken et al., in press) we obtained an L2 cost for text recall
but not for answering yes/no questions. The present study sought to find out to what
extent the L2 cost was a production cost (having to produce text in L2) rather than a
recall cost (having to recall information in L2).

The direction of results in Experiment 1B (L2-L2 < L2-L1) are consistent with the
previous finding that having to produce answers in L2 worsens performance, but the
effect was small (d = .12) and insignificant. At the same time, the data from Experiment 2 (and perhaps the non-significant trend in Experiment 1A) also indicate that having to study a text in L2 hampers performance (relative to studying the text in L1) for short open questions and for multiple-choice questions. This is different from our previous findings with yes/no questions, where we failed to find a significant difference between L1 and L2 even after one month.

In particular, the finding of worse performance with multiple choice questions in Experiment 2 was surprising to us, as we saw this as a recognition test (memory cues are given and participants do not have to write down the outcome nor organize the materials). Still, we saw a similar decrease in performance as in a test that required answering open questions. We argued that this could be due to a stricter time limit in Experiment 2, which is reflected in low mean scores, or to the fact that the multiple-choice questions of Experiment 2 required a deeper understanding than the true/false questions in our previous studies.

The findings of the present study resemble what was reported in a recent article by Roussel, Joulia, Tricot, and Sweller, (Roussel et al., 2017). These authors investigated language and content learning through texts in three conditions (L2 text, L2 text with L1-translation, and L1 text) 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. Roussel et al. (2017) tested the participants’ content knowledge with L1 questions probing cued recall. They observed that performance on the knowledge tests was best in the L1 condition, and worst in the L2 condition without translation support. However, the effect was not found in every experiment and seemed to depend on the proficiency level of the participants and on the topic (e.g. it was smaller for computer sciences, which contains more English terms in general).

Our findings, together with Roussel et al.’s (2017), illustrate the magnitude of the task we are confronted with to chart the L2 costs in memory for texts across a variety of educational situations. The following list of variables must play a role in the costs:
Type of text:
onfiction, short encyclopaedia-type entries, paragraphs from handbooks, journal articles;

Length of text:
going from paragraphs of 300 words to texts of over 10,000 words;

Proficiency of the participants:
ideally measured with a validated vocabulary test;

Time given for study and recall:
for instance, measured as minutes per 100 words;

Information about the language of the test at the time of study:
explicit or implicit instructions about the language of test, none at all, ...;

Language of the test:
L1 or L2; congruent with the study language or not;

Type of test:
free recall, open questions with long or short answers, multiple choice questions, yes/no questions;

Background (knowledge) of the participants:
computer science students versus law students for example.

Each study can only address a few cells of the complete mosaic. In addition, if we want useful information, each study must be properly powered (requiring a considerable investment of time and energy). Still, the enterprise is not impossible and we present Table 6 at the end of this chapter as a summary of what was already achieved. By augmenting the table in strategic ways, we can come to a meta-analytic review of the weights of the various factors involved, allowing us to give advice for a wide range of educational situations.

On a theoretical level, our studies are starting to provide some answers as well. The encoding-specificity principle does not seem to have a strong weight in L2 text studying. The only situation in which we expect it to play a discernible role is a situation in which participants answer questions about low-level aspects of the text. For instance, it could be that participants answer better to yes/no questions in a L2-L2 condition than in a L2-L1 condition. An interesting study in this respect was reported by Sahlin, Harding, and Seamon (2005). The authors presented Spanish-English participants with (auditory) lists of 10 English or 10 Spanish words, which the participants had to remember for a subsequent recognition test. In the recognition test, the words were presented in either the same language or in the other language amid distractors. Of the words presented in
the same language, 88% were recognized in L1 and 85% in L2. Of the words presented in translation only 13% were recognized in L2 and 17% in L1. Clearly, in this situation congruity between encoding and retrieval is critical. In related research, (Francis & Gutiérrez, 2012) reported that recognition memory for (visually presented) words was better in L2 than in L1 and attributed this to the existence of less memory interference in L2 between study and test on the one hand, and the fact that the encoding task was superficial on the other hand. The levels-of-processing theory (Craik & Lockhart, 1972) forms an interesting framework to interpret the findings and to search for other situations in which the encoding-specificity principle can be expected.

Does this mean that deep-level text memory is represented free of language, as Schank (1980) argued several decades ago? For the gist of the text, we endorse this point of view. The essence of a studied text is stored in a language-independent mental model. However, we are less sure about the details of the text. Maybe these are stored in a language-specific way, linked to the mental model? This could explain why answers to yes/no questions show the same forgetting curve in L2 as in L1. The critical test here will be to see whether such questions are answered better in a L2-L2 condition than in a L2-L1 condition (and possibly show a less steep forgetting curve).

The present results also make clear that L2 text production is not the only source of the L2 recall cost, as L1 recall of L2 text is also hampered to some extent (although L2 text production clearly contributes to the lower performance in L2-L2 free recall). A likely interpretation here is that a less mastered input language restricts the quality of the mental model that is constructed during the reading or studying of a text. Due to lower familiarity with L2 words or weaker conceptual representations of those words, the landscape model (dynamic, continuously updated mental representation of a text; van den Broek, Young, Tzeng, & Linderholm, 1999) is impoverished in L2 relative to L1, in particular when participants do not have enough time to study the text and/or when their proficiency scores are low relative to the difficulty of the text. As a consequence, the landscape model of an L2 text is not as active and robust as that of an L1 text, leading to poorer performance in recall or deep-level retrieval (for an earlier discussion of this interpretation, see Vander Beken et al., in press). Unknown L2 words are possibly lexically represented in this model (as a detail), but not conceptually, since the concept is not known. Similarly, in the landscape model, it is assumed that input concepts activate a cohort of related concepts and thereby enrich the mental model. Again, it is not
unreasonable to assume that L2 words fail to activate as many related concepts as L1 words do, and unknown words cannot activate any at all.

Two studies by Gablasova (2014, 2015) about learning specialized vocabulary from textbook reading are relevant in this respect. The studies compared how well participants acquire the meanings of technical words after reading and listening (simultaneously) to a text in L1 or L2. Though the vocabulary was acquired at a reasonable level in L2, the number of words and the number of meaning components recalled per word (respectively the breadth and depth) were significantly lower in L2 than in L1. Since all the technical words were cognates, meaning that they looked very similar in L1 and L2, the difference could hardly be related to the word forms themselves. In addition, the forgetting rate was higher in L2 on a delayed recall test after a week (Gablasova, 2014). The author 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. As a consequence, there is a “lower degree of elaboration” (Gablasova, 2014, p. 987) and in the longer term, more attrition is observed for those target words. This interpretation corresponds entirely with the arguments we have put forth.

Interestingly, Gablasova (2014) also mentioned the possibility that some knowledge which the participants cannot express, can be gained in a receptive manner. This passive knowledge cannot be recalled actively but can be recognised when strong memory cues are given. This may explain why the gaps in the mental model do not affect true/false questions, but impede recall and possibly also performance on more demanding multiple-choice questions (as observed in Experiment 2). Future research will have to disentangle these possibilities by testing both active and passive knowledge at the word and the text level. In addition, the large L2 recall cost that was found in the most complex task thus far, the free recall protocol (Vander Beken & Brysbaert, in press), could be examined further in a cross-lingual design and by means of qualitative research methods to see how much of the cost is due to the lack of L2 writing skill, and how much is due to the lack of actively retrievable information. This may also allow us investigate to what extent Gablasova’s claims can be extrapolated to the level of text retention.

The fact that the L2 cost varies from almost non-existent (Experiment 1) to $d = .8$ (Vander Beken & Brysbaert, in press) and depends on a considerable range of variables (Table 6) is frustrating for research purposes, because it shows the current gaps in our
knowledge. At the same time, the heterogeneity is very informative for educational purposes, and could be used by teachers to adapt their tests to L2 students. Certain types of tests (e.g., yes/no questions) can be used for inexperienced L2 students in situations where L2 is the only option, without harming the students’ performance (with the risk that the difference in depth of knowledge is not detected, which is a problem in its own right, as argued by Gablasova; 2015) Gradually the demands can be increased, as the L2 students learn more relevant words and build a richer mental model of the topic covered. Another way to improve performance may be to start with L1 for the basics and then switch to L2 when a firm foundation has been laid (Brysbaert & Dumoulin, 2007). The time allowed to study and to take a test is likely to moderate performance too (as suggested by Experiment 2).

All in all, our data indicate that the use of 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 may be wiser not to throw in students at the deep end, but to prepare them via a series of less demanding, intermediate goals. These consist of using questions that put less demand on uncued recall and English text production, and providing 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). To sum up: 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>
<thead>
<tr>
<th>Paper</th>
<th>Type of text</th>
<th>Length of text*</th>
<th>Language conditions</th>
<th>L1 and L2</th>
<th>Language factor</th>
<th>Type of test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen and Donin, 1997</td>
<td>Expository (handbook and procedural)</td>
<td>191-196 words</td>
<td>L1 - L1</td>
<td>L2 - L1</td>
<td>Chinese - English</td>
<td>within</td>
</tr>
<tr>
<td>Roussel et al., 2017</td>
<td>Expository (law text)</td>
<td>241 words</td>
<td>L1 - L1</td>
<td>L2 - L1</td>
<td>French - German</td>
<td>between</td>
</tr>
<tr>
<td>Roussel et al., 2017</td>
<td>Expository (law text)</td>
<td>241 words</td>
<td>L1 - L1 + L2</td>
<td>L1 - L1</td>
<td>French - English</td>
<td>between</td>
</tr>
<tr>
<td>Roussel et al., 2017</td>
<td>Expository (computer sciences)</td>
<td>324 words</td>
<td>L1 - L1</td>
<td>L2 - L1</td>
<td>French - English</td>
<td>between</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>Expository (encyclopedia-like)</td>
<td>258-279 words</td>
<td>L1 - L1</td>
<td>L2 - L2</td>
<td>Dutch - English</td>
<td>between</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Expository (encyclopedia-like)</td>
<td>258-279 words</td>
<td>L1 - L1</td>
<td>L2 - L2</td>
<td>Dutch - English</td>
<td>Within (delay = between)</td>
</tr>
<tr>
<td>Chapter 4 Experiment 1A</td>
<td>Expository (handbook)</td>
<td>488-517 words</td>
<td>L1 - L1</td>
<td>L2 - L1</td>
<td>Dutch - English</td>
<td>within</td>
</tr>
<tr>
<td>Chapter 4 Experiment 1B</td>
<td>Expository (handbook)</td>
<td>488-517 words</td>
<td>L2 - L1</td>
<td>L2 - L2</td>
<td>Dutch - English</td>
<td>within</td>
</tr>
<tr>
<td>Chapter 4 Experiment 2</td>
<td>Academic article (part of it)</td>
<td>1749 words</td>
<td>L2 - L1</td>
<td>L1 - L1</td>
<td>Dutch - English</td>
<td>between</td>
</tr>
</tbody>
</table>

* 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.
* Based on English versions, even if there is an experiment which includes another non-native language
The means are not reported, but more time was spent in L2 reading: F(2,32)=86.99. **This does not take into account the translated text in the translation condition (which ought to result in about twice the number of words to read in the same time).**

The authors report to have split up the participants in a lower and higher proficiency group, using a Michigan Test of English Language Proficiency score of 79 as threshold. However, they do not report the means of those groups. A cut score of 64 is reported to accord to a B2/C1 CEF level, though this comparison is not validated, so we must be cautious in interpreting this (see [http://cambridgemichigan.org/institutions/products-services/placement-progress/mtelp-series/levels-scoring/](http://cambridgemichigan.org/institutions/products-services/placement-progress/mtelp-series/levels-scoring/)). The authors reported the groups as low-intermediate to intermediate versus high-intermediate to high.
References


STUDYING TEXTS IN A NON-NATIVE LANGUAGE:
THE L2 RECALL COST
IS NOT JUST A PRODUCTION EFFECT

Heleen Vander Beken & Marc Brysbaert
ABSTRACT

Bilingualism research has long tried to answer the question whether meaning is represented at a shared, conceptual level or in a language-dependent manner, but most studies have looked at the processing of words or sentences in isolation. With internationalisation of education, it is becoming even more important to understand language effects at higher levels such as the studying of texts in a non-native language (L2). In previous studies, we found a large recall cost for short, expository texts in L2 compared to L1, but no such disadvantage for a true/false recognition test (Vander Beken & Brysbaert, in press). The finding that no recognition effect occurred even on the long term seemed to suggest that production in L2 was hampered, but not encoding. Nevertheless, other tasks such as cued recall did yield some effects (Chapter 4). This study compares recall of an L1 text and an L2 text when the recall protocol is written in the native language. In addition, it is compared how well an L2 text is recalled when the recall protocol is written in L1 versus in L2. No significant effect is observed for the language of recall, while an L2-cost is found for the language of encoding (Cohen’s d = .6). This suggests that the encoding process is responsible for most of the recall cost and provides evidence for language-independent representation of meaning.
Introduction

English is increasingly used as a lingua franca and, with rapid internationalisation of higher education, as a medium of instruction. In addition, mobility of students results in many situations of non-native language education. Hence, over the last couple of years, policy makers and scientists have intensively been looking into the benefits and the risks of using a non-native language for educational purposes.

Next to ideological and financial arguments and objections, the most important question is whether the same quantity and quality of knowledge is gained in a weaker language context as in the native language. From an educational point of view, this boils down to the question whether a non-native language is a feasible challenge that might even optimise learning (a “desirable difficulty”, Metcalfe, 2011) or rather an extra obstacle, raising the cognitive load in an already difficult task of processing lots of new, specialised information. A straightforward way to answer this question is to compare performance in exactly the same learning task in a native (L1) and a non-native (L2) language, in this case English. This experimental take allows us to understand the cognitive consequences of foreign language education and improves our understanding of bilingual memory for learnt information in general.

Two frameworks from memory and language research predict contradictory effects for studying in L2. First, if memory is language-independent, we could state that language only affects performance by mediating the quality of the input into memory and the quality of the memory output (several frameworks can account for this, such as the RESOURCE HYPOTHESIS (Sandoval, Gollan, Ferreira, & Salmon, 2010) and the WEAKEr-LINKS HYPOTHESIS (Finkbeiner, 2002)). A language-independent view means that the relationship between the language in which one learns and the language in which one recalls information is unimportant. On the other hand, the memory literature has reported context-dependency or congruency effects with better performance when the context of encoding and retrieval are similar. This is called the ENCODING-SPECIFICITY PRINCIPLE (Tulving & Thomson, 1973), and there are some instances of this effect for language (e.g. Marian & Fausey, 2006; Marian & Neisser, 2000). The first framework, the widely-spread language-independent view, hypothesizes that performance possibly
becomes worse every time a weaker language is involved. The second idea, based on the encoding-specificity principle, predicts lower performance on cross-lingual tasks (e.g. L1 recall of L2 text), and perhaps even the best performance in L2-L2 conditions, since L2 memory traces are more unique and suffer less from interference of new information (Francis & Gutiérrez, 2012). Of course, a combination is possible if memory is indeed language-dependent and benefits from congruency, but is still affected by the lower quality of L2 in- or output. If an L2-disadvantage arises, as one would intuitively expect, the question remains at which proficiency level it occurs, what size the effect has and which processes are behind it.

Recently, a number of experimental paradigms have led to diverse and inconsistent findings. Our own studies have shown that recognition memory, tested with true/false judgements, is not hampered in congruent L2 versus L1 conditions, even on the long term (Vander Beken & Brysbaert, in press; Vander Beken, Woumans, & Brysbaert, in press). This was also true for cued recall in another study (Chapter 4, Experiment 1), in which we used open and closed questions. In that study, there was only a small and insignificant trend in the direction L1-L1 > L2-L1 > L2-L2. This finding corresponds to observations reported by Chen and Donin (1997), who tested online and offline oral free recall in the same three language conditions with Chinese as L1 and English as L2. No significant difference was found, but there was a trend of lower L2 recall of an L2 text than L1 recall, which might have been insignificant due to low power in the study.

In contrast, in Chapter 4 (Experiment 2) we also obtained a considerable L2 cost with a recognition task, consisting of multiple choice questions. This unexpected recognition cost may have been because the texts were considerably longer and our time limit was too short. Perhaps students were not able to finish studying in L2 (reading and studying rate in university students in L2 is about 18% slower than in L1; see Dirix et al., submitted). Roussel, Joulia, Tricot, & Sweller (2017) also reported significantly lower scores for law students on an L1 content test (i.e., cued recall) after learning texts in L2 compared to L1, but they did not replicate this effect with computer science students. Furthermore, Gablasova (2014, 2015) reported an effect on the depth and breadth of new technical vocabulary knowledge that was learnt through the reading of a text with definitions of those words, with less knowledge to start with and more forgetting in L2.
(the test language was not taken into account but counterbalanced, half in L1 and half in L2).

The clearest evidence for an L2 cost so far was reported by Vander Beken and Brysbaert (in press). They used a free recall task – in which participants had to write down as much as they could remember from the learned information. Performance was substantially worse in L2 compared to L1, with an effect size of Cohen’s $d = .8$. A similar recall cost was reported in a small-scale study by Connor (1984): L1-speakers recalled significantly more propositions from a text than L2-speakers of English (at least for subordinate ideas; no difference was found for higher-level ideas).

There are two possible explanations for the large cost in free recall, compared to the rather small or non-existent costs in recognition and cued recall. On the one hand, L2 production may be impaired to such an extent that the learned knowledge cannot fully be expressed (despite the fact that language mistakes were not punished in the scoring). On the other hand, it is also possible that some knowledge is gained receptively, and thus can be recognised in a true/false test but not actively retrieved and reproduced (this can be considered MARGINAL KNOWLEDGE, see Cantor, Eslick, Marsh, Bjork, & Bjork, 2014; Vander Beken & Brysbaert, in press; Gablasova, 2014). The more superficial processing in L2 is in line with the LEVELS-OF-PROCESSING EFFECT (Craik and Lockhart, 1972; see also Francis & Gutiérrez, 2012). The observation of an L2 effect on multiple choice-questions also points in the direction of a weaker mental model of the text, rather than a mere production issue. A less expected finding, however, is that no effect of L2 was found on a long-term test of true/false judgements (after one week or one month). Weaker memory traces in L2 would arguably yield a steeper forgetting curve.

In summary, performance in L2 is usually worse than in L1. It seems like the L2-effect is mediated by several factors, such as time spent on the task, the difficulty level of both the text and the test, the familiarity of the topic, proficiency level, the number of cues available during the task, etc. (see Chapter 4 for a first overview of these factors). The difference is not significant for yes/no recognition tests, significant for some cued recall tests (and insignificant for others), larger for a rather demanding multiple choice test, and most pronounced for a free recall task. We propose two elements to account for these findings: an L2 production cost and a more superficial organisation in memory for
information studied in L2. To disentangle these accounts, the production cost in the free recall task (in which the largest effect was reported) should be tested directly.

This leads us to the issue of written recall protocols in native and non-native languages, a matter which has perhaps received more attention in research than the relationship between language and memory itself. For example, qualitative research showed that strategies used in L1 and L2 text recall are comparable (Appel & Lantolf, 1994) and that L2 recall is affected by factors such topic interest and prior knowledge (Joh, 2006). Quantitative studies showed a disadvantage in the number of ideas recalled in L2 (see above, Connor, 1984), albeit in a limited sample size. Furthermore, like reading in L2 is impaired and slower than in L1, the same holds for writing in a weaker language. Chenoweth and Hayes (2001) experimentally showed that language experience impacts the number of words written per minute in a recall task. In their population, the ratio of writing speed in L2 versus L1 was 0.66, and the process of creating the L1 text was more fluent, with less revisions and longer “bursts” in which language was created. A more recent study in this respect was reported by Chang (2011), who investigated the relation between time and difficulty level in an L1 recall task of an L2 text. Performance suffered from time constraints, and this effect was mediated by the difficulty of the text passage (time was measured over studying and writing the recall). These findings once more point towards the contribution of production issues in L2 recall difficulties. The next step to understand the contribution is to zoom in on the specific recall effect in a cross-lingual situation.

The current study uses the same materials from the study in which the L2 recall cost was found in unilingual and cross-lingual conditions. Language is tested as a within-subject variable in two separate experiments (comparing L1-L1 with L2-L1 and L2-L1 with L2-L2). In both experiments, participants receive some vocabulary support and are presented with a free recall test first and true/false test subsequently. The reason for giving vocabulary support is that we cannot expect subjects to perform well in an L2-to-L1 condition if they do not know some L2 words and their translations. This might reduce the L2 disadvantage compared to an earlier study (Vander Beken & Brysbaert, in press), but at the same time, we can observe whether the L2 recall cost remains when subjects get some support. For the same reason, explicit instructions were given about the language condition and the type of tests at the start of studying the texts. This way,
subjects could choose the optimal strategy to prepare for the test (and, for example, realise they needed the vocabulary support). In the previous study, subjects did not receive vocabulary support or explicit instructions about the type of test.

We expect to find a difference on the recall test in both experiments, but not on the recognition test. If a difference is found for L1 recall of an L1 text versus an L2 text, this provides evidence that the encoding language has an effect on memory for texts (or the mental model of a text). If a difference is found for L1 recall versus L2 recall of an L2 text, this means that the language of recall has an effect on memory performance and that the large L2 recall cost reported in an earlier study (Vander Beken & Brysbaert, in press) is (also) caused by production difficulties. If any difference is found for the recognition tests, we predict the cross-lingual condition to be the lowest (since language-congruency effects have been reported on recognition tasks).

Method

Participants

Since we are working with a repeated-measures design, we need a group of 52 participants to have 80% chance to detect a main effect of a within-subjects variable with $d = .4$, at least when each participant reads only one text per condition (Brysbaert & Stevens, 2018). 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, and it was required that 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. In some of the university courses English handbooks are used as well, even though the teaching happens in Dutch.
In Experiment 1, 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 2, 60 students were tested. After exclusion of a student that had participated in a previous study (despite selection criteria), 59 students remained\(^1\).

Materials

Texts

We used the materials from a previous study (Vander Beken & Brysbaert, in press), which are adapted versions of two short, English texts from a study of Roediger and Karpicke (2006). Those texts were translated to Dutch and matched on word frequency measures between languages. In the present study, they 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

Since this study contains an experiment in which students study a text in L2 and are asked to express their knowledge in L1, we decided to add vocabulary support. If participants are not aware of the translation of a word, they cannot express their memory of it. Therefore, a number of words was selected for each text based on (a) a post-test about vocabulary from the Sea otters text that was administered after an experiment using the same texts (Vander Beken et al., in press) and (b) the word frequencies. In the

\(^1\) For data transparency, we must add that an additional 26 participants were tested. When testing the 54th participant, a programming error was discovered in the true/false judgement 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 was discarded and replaced by 26 new testing session. 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.
vocabulary post-test, participants were asked to express their understanding of words which we suspected to be problematic by providing an explanation, definition, or translation (whichever they preferred). Words that were not known by at least 80% of the participants were included in the vocabulary support in the present study, together with some low-frequent words with a Zipf-value lower than 3.6 (for The Sun, there was no post-test, so only the frequency criterion was used and the selection was elaborated using common sense). If these low-frequent words were interlingual cognates, morphological derivations of words that are assumedly known, or had received a definition in the text itself, they were excluded. The remaining words were presented in a list together with the texts, containing a translation or explanation of the words. For similarity, the Dutch texts received a list in which the same items (that is, the translation equivalents of the English list) were presented with a synonym or hypernym (a category label). Although these Dutch words perhaps did not need clarification in all cases, the presentation of a word list might prime these concepts and affect memory on the one hand, or distract attention from the text on the other hand. Therefore, we constructed the Dutch list, so that vocabulary support was present in all conditions. For the text about the Sun, the word list contained 10 items, for Sea otters, it contained 20 items (see Appendix A.2.1).

**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 taken from the same study (Vander Beken & Brysbaert, in press) and adapted to some extent. 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.*

This way, participants were not asked to literally reproduce the text, but to produce the ideas and to add details when possible.

Roediger and Karpicke (2006) divided their texts into 30 ideas or propositions that had to be reproduced. This list (with adaptations analogous to the text adaptations) was
used as a scoring form for the free recall tests in English, and a Dutch version was created (see Vander Beken & Brysbaert, in press).

From those propositions, an equal number of true/false questions was created, complemented with some additional questions of different types (inferences and false memory questions), adding up to 46 questions (see Vander Beken & Brysbaert, in press). Questions that appeared to be passage-independent (meaning they can be answered based on prior knowledge without reading the text passage) in a pilot study and were excluded from all analysis before, were now replaced by new questions. 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 or on the top of the screen for every test in the language of the test. Answer option were “True”, “False”, or “I don’t know”. The opt-out option was added in a long-term recognition study (Vander Beken et al., in press) to avoid a large guessing effect and has been kept here. All tests were administered on LimeSurvey. The texts and the tests can be found in Appendix A.

**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, in press). Some propositions were additionally split up in two because we had the experience participants frequently only recalled half of these ideas. The mean was calculated for all of those split up ideas, so that an average score could be taken in a similar way to the first study. 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. Once more, 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. There were categories for correct and incorrect/incomplete answers as in the previous studies (scored 0, the previous study showed that a 0.5 option for single propositions was barely used and differed between raters). In addition, three other categories were added because of the cross-lingual condition: a category for correct English answers on a Dutch recall form,
incorrect answers that could be traced back to the non-target language (e.g. when someone mistranslates an idea) and misinterpretations that must have occurred in the target language. To calculate the total memory score for every test, correct answers in the correct language and correct answers in the wrong language were given a score of 1, all other categories were transformed to a score of 0 (the previous study showed that a 0.5 option for semi-correctly recalled ideas was barely used and differed between raters).

In order to determine interrater reliability for nominal data rated by two raters, we calculated Gwet’s AC₁ (Gwet, 2008). This has the advantage over the more commonly used Cohen’s kappa that it is robust in cases of high agreement or prevalence (Wongpakaran, Wongpakaran, Wedding, & Gwet, 2013). In this dataset, certain categories (correct answers and incorrect/missing answers) make up most of the ratings. Their prevalence could affect the reliability score (Hallgren, 2012) when using 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 1, AC = 0.88 (SE = 0.009) and in experiment 2, the AC = 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 ratings of the first rater. The counts of the categories in half of the protocols (rated by both raters) are displayed in Table 1.

### Table 1. Counts of categories in the free recall ratings that were rated by two raters (half of all protocols, 1731 observations for Experiment 1, 1745 observations for Experiment 2).

<table>
<thead>
<tr>
<th>Categories</th>
<th>Exp1-R1</th>
<th>Exp1-R2</th>
<th>Exp2-R1</th>
<th>Exp2-R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (incorrect)</td>
<td>696</td>
<td>732</td>
<td>783</td>
<td>785</td>
</tr>
<tr>
<td>1 (correct)</td>
<td>981</td>
<td>956</td>
<td>918</td>
<td>935</td>
</tr>
<tr>
<td>2 (correct in other language)</td>
<td>19</td>
<td>25</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>3 (incorrect derivation from other language)</td>
<td>27</td>
<td>21</td>
<td>27</td>
<td>23</td>
</tr>
<tr>
<td>4 (incorrect within language)</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>
To calculate memory scores, the ratings were transformed into a dichotomous score for correctness (with 2 coded as correct and 3 and 4 as incorrect). There was one specific exception for the recall protocols of *The Sun*. Three ideas in that text contain the numeric term “billion” \(10^{9}\), which is an interlingual homograph: in Dutch it translates to “miljard”, while “biljoen” means a trillion \(10^{12}\). 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. Nevertheless, on an actual exam, they would be considered wrong. The average of the dichotomous scores was then taken over all propositions, expressed in percentages. For comparison with the previous study (Vander Beken & Brysbaert, in press) and the original study (Roediger & Karpicke, 2006), we also calculated a partial score based on the 30 ideas by first averaging all ideas that we have split up to a score on 1, and then averaging those 30 to a partial total score.

The true/false tests were scored dichotomously (correct/incorrect). “I don’t know”-answers were scored as incorrect. The questions that were excluded from the previous study (Vander Beken & Brysbaert, in press) were replaced in this study, but not included in the main analysis (since they would not be comparable). A separate analysis was done with the new questions included, because we think they will be interesting for future research.

Control variables

We collected proficiency and working memory data for all participants. Working memory capacity was measured with the automated operation span task (Unsworth, Heitz, Schrock, & Engle, 2005), administered in E-Prime 2.0.10. Table 2 displays the means and standard deviations of the operation span task and of the following proficiency measures.

*Objective L1 proficiency tests*

L1 receptive vocabulary knowledge was measured with a semantic vocabulary test in a multiple choice format with four answer alternative. In addition, participants received a Dutch 46-item spelling test in which they had to type out words of various spelling difficulties that were read aloud (developed from a previous 92-item version, Vander
Beken & Brysbaert, in press). In Experiment 2, the full 92-item version was administered since the shortened version used in Experiment 1 appeared to be too easy, despite our effort to select balanced items. The spelling tests were added here since they are a measure of written productive proficiency, which plays a large role in the free recall task. The Dutch proficiency tests are available in Appendix C.

**Objective L2 proficiency tests**

L2 proficiency was measured with the English LexTALE test of vocabulary knowledge for advanced learners of English (Lemhöfer & Broersma, 2012). Next, a 45-item English spelling test was given, similar to the Dutch spelling test (developed from a previous 92-item version, Vander Beken & Brysbaert, in press). For Experiment 2, the full 92-item version was used, in accordance with the Dutch spelling test (see Appendix C). The spelling tests were added here since they are a measure of written productive proficiency, which plays a large role in the free recall task. In addition, the first part (40 items) of the Oxford Quick Placement Test (QPT; 2001) was also administered, which is considered a measure of general proficiency consisting of multiple choice items of vocabulary and sentence comprehension and grammar (verb use, part of speech regulations, ...).

Results indicate that L2 proficiency (mean score for the LexTALE was 75.56 in Experiment 1 and 74.12 in Experiment 2) is comparable to previous chapters with similar participant groups, and thus much lower than L1 proficiency. English and Dutch vocabulary scores and working memory capacity are also similar to those studies. Table 5 and 6 (see Appendix to this chapter) contain the reliabilities and correlations between the measures.
Table 2. Mean scores on the various proficiency and intelligence tests of the participants in Experiments 1 and 2 (standard deviations between brackets).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Exp. 1 (N = 56)</th>
<th>Exp. 2 (N = 59)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>39F/17M</td>
<td>46F/13M</td>
</tr>
<tr>
<td>Age</td>
<td>21.71 (3.21)</td>
<td>20.37 (2.23)</td>
</tr>
<tr>
<td>English LexTALE (max = 100)</td>
<td>75.56 (11.82)</td>
<td>74.12 (10.66)</td>
</tr>
<tr>
<td>English Spelling (%)</td>
<td>76.23 (11.9)*</td>
<td>54.99 (15.01)*</td>
</tr>
<tr>
<td>Oxford Quick Placement test (max = 40)</td>
<td>32.98 (3.81)</td>
<td>32.78 (4.04)</td>
</tr>
<tr>
<td>Dutch vocabulary MC (max = 75)</td>
<td>49.00 (8.47)</td>
<td>47.32 (6.77)</td>
</tr>
<tr>
<td>Dutch Spelling</td>
<td>94.72 (5.03)*</td>
<td>81.11 (7.30)*</td>
</tr>
<tr>
<td>Operation Span (WM) (max = 75)</td>
<td>60.39 (10.96)</td>
<td>59.69 (10.52)</td>
</tr>
</tbody>
</table>

*Note: the spelling scores in Experiments 1 and 2 are based on a different test (twice as long in Experiment 2, hence the difference in scores).

Subjective assessment of language proficiency

The participants’ language background 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). The self-ratings of L1 and L2 proficiency are displayed in Table 3.

Motivation and Text-related Questionnaires

After the free recall and true/false tests, participants completed a questionnaire asking about their general motivation and attitude towards testing and towards reading in L1 and L2. The questionnaire also contained questions about their text-specific experiences (checking for prior knowledge about the texts, the perceived difficulties, and how interesting the texts were). The questionnaire was presented in Dutch to all participants, using 7-point Likert scales. These self-ratings of motivation are displayed in Table 3 and the questionnaire was added in Appendix D.
Table 3. Mean scores of the language groups on the self-ratings included in the questionnaire (standard deviations between brackets).

<table>
<thead>
<tr>
<th>Self-ratings</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General motivation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test importance (7)</td>
<td>5.48 (1.22)</td>
<td>5.42 (0.99)</td>
</tr>
<tr>
<td>Performance vs. peers (7)</td>
<td>4.16 (0.76)</td>
<td>4.00 (0.81)</td>
</tr>
<tr>
<td><strong>Dutch academic reading</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitude (7)*</td>
<td>5.2 (1.12)</td>
<td>5.17 (0.83)</td>
</tr>
<tr>
<td>Intrinsic motivation (7)*</td>
<td>4.92 (0.99)</td>
<td>4.92 (0.8)</td>
</tr>
<tr>
<td>Total motivation (7)*</td>
<td>5.05 (0.92)</td>
<td>5.14 (0.65)</td>
</tr>
<tr>
<td><strong>English academic reading</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitude (7)*</td>
<td>5.56 (1.04)</td>
<td>4.72 (0.94)</td>
</tr>
<tr>
<td>Intrinsic motivation (7)*</td>
<td>4.85 (0.99)</td>
<td>4.48 (0.89)</td>
</tr>
<tr>
<td>Total motivation (7)*</td>
<td>4.67 (1.02)</td>
<td>4.62 (0.79)</td>
</tr>
<tr>
<td>Opinion about use of EMI (7)*</td>
<td>6.07 (0.97)</td>
<td>6.00 (0.79)</td>
</tr>
<tr>
<td><strong>Dutch language skill</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading (10)</td>
<td>9.52 (0.69)</td>
<td>9.29 (0.86)</td>
</tr>
<tr>
<td>Writing (10)</td>
<td>9.07 (0.97)</td>
<td>8.58 (1.16)</td>
</tr>
<tr>
<td>Proficiency (10)*</td>
<td>9.43 (0.65)</td>
<td>9.15 (0.74)</td>
</tr>
<tr>
<td><strong>English language skill</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading (10)</td>
<td>7.96 (1.21)</td>
<td>7.76 (1.16)</td>
</tr>
<tr>
<td>Writing (10)</td>
<td>7.13 (1.28)</td>
<td>8.58 (1.50)</td>
</tr>
<tr>
<td>Proficiency (10)*</td>
<td>7.75 (0.87)</td>
<td>7.38 (1.04)</td>
</tr>
</tbody>
</table>

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

Since we had two texts at our disposal from the Vander Beken and Brysbaert (in press) study, we split up the study in two smaller experiments with a repeated-measure design. In Experiment 1 (N = 56) participants received one text in English and one in Dutch, but received Dutch tests for both (language of text as repeated measure). The language and order of the texts were counterbalanced between participants, to avoid text-specific effects. In Experiment 2 (N = 59), participants received both texts in English, but received one test in Dutch and one in English (language of test as repeated measure). The order of the texts and language of the tests was counterbalanced between participants.

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 (2 x 2 factorial design).

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 and detailed questions) afterwards, with a time limit of 7 minutes for every part. In experiment 1, they were informed that one text would be in English and one in Dutch, but that all tests were in Dutch. In experiment 2, participants 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. This way, they had all necessary information to prepare the best way they could, as if it were for a real exam. Texts and recall tests were presented on paper; the true/false test was administered in LimeSurvey.

After the text study phase, the spelling tests were administered in group. The words were read aloud by the same experimenter in every group and repeated when anyone asked. For the English version, both British and American spelling and homophones were considered correct, which was also mentioned in the instructions. Participants typed out the words in an Excel file.
Afterwards, participants completed the operation span task individually and continued by filling in the various questionnaires and vocabulary tests at their own pace. The experiment took approximately two hours in total, plus or minus 15 minutes due to individual or group differences in speed.

Results

All data are available at https://osf.io/p5b3y/ (Open Science Framework). In Table 4, an overview of memory scores is provided.

<table>
<thead>
<tr>
<th>Memory tests</th>
<th>L1-L1</th>
<th>L2-L1</th>
<th>L2-L1</th>
<th>L2-L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>True/false (Chapter 2 scoring)</td>
<td>69.23</td>
<td>65.88</td>
<td>66.83</td>
<td>68.42</td>
</tr>
<tr>
<td>True/false (adapted scoring)</td>
<td>69.14</td>
<td>66.15</td>
<td>66.85</td>
<td>66.98</td>
</tr>
<tr>
<td>True/false (Vander Beken &amp; Brysbaert, in press)</td>
<td>80.9</td>
<td>80.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free recall (Chapter 2 scoring)</td>
<td>62.48</td>
<td>53.99*</td>
<td>49.89</td>
<td>51.64</td>
</tr>
<tr>
<td>Free recall (adapted scoring)</td>
<td>63.87</td>
<td>54.88*</td>
<td>51.66</td>
<td>53.30</td>
</tr>
<tr>
<td>Free recall (Vander Beken &amp; Brysbaert, in press)</td>
<td>56.30</td>
<td></td>
<td>44.00</td>
<td></td>
</tr>
</tbody>
</table>

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

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. First, the partial total scores that are comparable to Vander Beken & Brysbaert (in press, Chapter 2) are considered, followed by a new calculation including new questions2.

2 We included this new calculation because passage-independent questions were replaced. Including new items allowed us to test every idea present in the text and to see whether this makes any difference. In other words: the new calculations are more complete, but the old version is more comparable. Therefore, both are included.
Experiment 1 (Chapter 2 scoring):

Based on the percentages correct, the mean recognition score was 69.23% (SD = 11.15) in the L1-L1 condition and 65.88% (SD = 12.56) in the L2-L1 condition. Since the data were normally distributed, a two-sided paired t-test was used for analysis. A statistically significant difference was found (t(55) = 2.14, p = .04). A post-hoc calculation of effect size results in Cohen’s d = .29 (a small effect), with a confidence interval between -.09 and .67. Next to hypothesis testing, we also ran Bayesian analyses to assess the amount of evidence for or against the null hypothesis. The Bayes factor B10 was 1.2, which denotes anecdotal evidence for the alternative hypothesis. Since both the confidence interval and the Bayes’ factor do not point towards an effect and the t-tests is borderline significant, this can be considered a very small or non-existent effect.

Experiment 2 (Chapter 2 scoring):

Mean recognition was 66.83% (SD = 13.82) in the L2-L1 condition and 68.24% (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 B10 was 0.84 here, which denotes anecdotal evidence for the null hypothesis.

Comparison to previous results:

By juxtaposing the scores in the cross-lingual condition in Experiment 1 and 2, we can conclude that both experimental groups perform at a similar level (65.88% vs. 66.83%).

The score of the unilingual conditions in Experiment 1 and 2 can be compared to those of Vander Beken and Brysbaert (in press) to check whether the results are replicated (despite being in different groups).

The mean score in the L1-L1 condition (experiment 1) is 69.23% (SD = 11.15) and in the L2-L2 condition (experiment 2) is 68.24% (SD = 13.50). In the same conditions of Vander Beken and Brysbaert (in press), these were 80.9% (SD = 11.8) and 80.1% (SD = 8.7). Hence, the scores in the present study are lower than in the previous study, probably due to a diminished guessing factor, but the language conditions are very close...
to each other once again. In the latter sense, the results from the previous study are replicated.

**Including additional questions (new calculation)**

When we include the replaced questions (to reach optimal variance with a maximum of questions), findings are similar. The mean recognition score was 69.14% (SD = 12.56) in the L1-L1 condition and 66.15% (SD = 11.93) in the L2-L1 condition. Since the data were normally distributed, a two-sided paired t-test was used for analysis. No significant difference was found (t(55) = 1.95, p = .06). The Bayes factor was 0.85 which denotes anecdotal evidence in the opposite direction, that is, for the null hypothesis.

For experiment 2, the mean recognition score was 66.85% (SD = 13.32) in the L2-L1 condition and 66.98% (SD = 12.85) 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) = -1.03, p = .31). In this case, the means indicate that there is barely any difference. The Bayes factor B_{10} was 0.25 here, which denotes moderate evidence for the null hypothesis.

**Memory performance on the free recall tests**

Firstly, the total scores based on the same 30 ideas as in the previous study and Roediger and Karpicke’s work are considered. Next, these are compared to our adapted recall key, since we believe this provides a better estimation of the total information that is recalled.

**Experiment 1 (Chapter 2 scoring):**

Based on the percentages correct, the mean recall score was 62.48% (SD = 12.89) in the L1-L1 and 53.99% (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. The confidence
interval of the effect size ranged from 0.23 to 1.00. The Bayes factor $B_{10}$ was 746.57 here, which denotes extreme evidence for the alternative hypothesis.

**Experiment 2 (Chapter 2 scoring)**

Based on the percentages correct, the mean recall score was 49.89% (SD = 14.41) in the L2-L1 and 51.64% (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 here, which denotes moderate evidence for the null hypothesis.

**Comparison to previous results:**

By juxtaposing the scores in the cross-lingual condition in Experiment 1 and 2, we can conclude that both experimental groups perform at a comparable level, though the first group performed slightly better (53.99% vs. 49.89%, mean difference of about one idea).

Using this calculation of the total score, the unilingual conditions in experiment 1 and 2 can be compared with Vander Beken and Brysbaet (in press) to check whether the results are replicated (despite being in different groups).

The mean score in the L1-L1 condition (experiment 1) is 62.48% and in the L2-L2 condition is 49.89%. These scores are slightly higher in comparison to the 2017 study (with scores of respectively 56.3 and 44) but with a similar difference. It must be remembered that in the present experiment we informed the participants about the language of the upcoming test and also gave them vocabulary support.

**Adapted recall key (new calculations):**

Based on our adapted recall key, the pattern of results for experiment 1 is exactly the same, with mean scores of 63.87% (SD = 12.39) in the L1-L1 condition and 54.88% (SD = 15.05) in the L2-L1 condition. A one-sided t-test was used (since we expect a higher score in the native language condition). Performance was significantly better in the L1-L1
condition \((t(55) = 4.61, p < .001)\). Bayes factor \(B_{10}\) was 819.06 here, which denotes extreme evidence for the alternative hypothesis.

For experiment 2, the pattern of results is also exactly the same, with a mean recall score of 51.66\% (SD = 14.61) in the L2-L1 condition and 53.30\% (SD = 14.00) in the L2-L2 condition. No significant difference was found on a two-sided t-test \((t(58) = -0.68, p = .5)\). The Bayes factor \(B_{10}\) was 0.18 here, which denotes moderate evidence for the null hypothesis.

**Incomplete protocols**

In addition, we took note of recall protocols that were incomplete when we suspected this was due to lack of time. In some cases, protocols were clearly unfinished, so we distinguished between these obvious cases and more speculative cases. In Experiment 1, we observed 7 definite cases of time shortage and 9 probable cases, in Experiment 2 respectively 7 and 6. Since ideas can be missing even when someone finished his protocol, there are probably more actual instances of time shortage than we could observe, which might have affected performance (and possibly enlarges the language effect).

![Figure 1. Memory scores in percentages per language condition and test type in Experiment 1 (all tests in L1) with 95% confidence intervals.](image-url)
Discussion

In the present paper we tested to which extent the free recall cost in a second language 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 as in a previous study (Vander Beken & Brysbaert, in press) in which a recall cost of $d = .8$ was found in an L2-L2 condition compared to L1-L1, while no cost at all was found in a recognition test.

Two possible explanations were given in that study. The first was centred on a production deficit; the second assumed that the mental model of the text was weaker because processing is more superficial in L2. This would lead to unhampered recognition.

Figure 2. Memory scores in percentages per language condition and test type in Experiment 2 (all texts in L2) with 95% confidence intervals.
memory but worse recall. Evidence against the latter hypothesis was found in (1) the fact that recognition memory did not decay more in L2 than in L1 (this should be the case with weaker memory traces; Vander Beken, Woumans & Brysbaert, in press), (2) experiments with open questions that did not yield significant differences between L2-L1 and L1-L1 conditions (Chapter 4; Chen & Donin, 1997). Still, some evidence raised doubts about the production account as well. For example, no significant disadvantage was found on open questions in an L2-L2 versus L2-L1 condition (Chapter 4; Chen & Donin, 1997), while in other situations, a disadvantage was found for native tests of an L2 text versus an L1 text, in which production does not play a role (multiple choice questions in Vander Beken, De Bruyne, & Brysbaert, submitted; Roussel et al., 2017). The most reasonable assumption is that the actual cost depends on the task requirements (amount of production needed), the time given for the task, the difficulty of the text and test, etc (see discussion in Chapter 4). In addition, a cross-lingual test of the original recall task with the large L2 cost can provide more information on where the cost is located.

First, the results on the true/false test once again confirm that an L2-disadvantage does not occur in all situations. For tests of recognition memory that only require confirmation or rejection of propositions, participants perform at a similar level in unilingual conditions in L1 and L2. There is an indication that a cross-lingual test of recognition results in worse scores, but this effect is only found in comparison to the native language condition and is not supported by the Bayesian analysis and the confidence interval. Hence, it could be argued that recognition tests tap into superficial knowledge of the text that has been read, and that it becomes more challenging when concepts need to be accessed to respond to the statements. Still, evidence is limited and no large disadvantage has been found in cross-lingual conditions. Moreover, performance barely benefits from a test in the native language, if it benefits at all.

Second, a very important conclusion can be drawn about the L2 recall cost: production difficulties do not account for the large disadvantage in L2 recall of L2-text. The same result is found when participant can write their recall in L1 as in L2. This finding is more in line with a weaker mental model: deep encoding or storage of information in L2 compared to L1 is hampered. In addition, the effect size found here (d ≈ .6) is not much smaller than that found in Vander Beken and Brysbaert (in press: d ≈
As a caveat, we should mention that an old study (Lee, 1986) reports a different conclusion. This study is framed in terms of reading comprehension rather than memory for text. The ambiguity about reading comprehension research is that reading comprehension can be tested with or without text accessibility at the time of the test (Schaffner & Schiefel, 2013). Needless to say, the second situation is very similar to our memory test. Lee (1986) compared reading comprehension in L2-L1 and L2-L2 conditions with a free recall test very similar to ours, and found better performance in L2-L1 (M = 9.0 ideas recalled) than in L2-L2 (M = 7.6). It is not clear what caused the difference between Lee's (1986) results and ours. Four factors may be involved. First, it may be a consequence of the vocabulary support and instructions given. Second, Lee's (1986) English-Spanish bilinguals might have had a different L2 proficiency than our participants. Lee tested foreign language learners of Spanish at four different levels, namely the first up to the fourth semester of Spanish university course (first and second year). So the range of proficiency is assumedly broader than in our homogenous group, but proficiency level was not measured, nor was a minimally required proficiency level mentioned for the course levels. Third, the scores of Lee (1986) were much lower than ours. Participants only recalled 6-10 of the 34 ideas, whereas our participants recalled about 60% of the ideas presented (note that the L2 cost for cued recall in Chapter 4 Experiment 2 was also found in very low mean scores). Finally, no information is given by Lee (1986) about the scoring. In particular, it is possible that the lower scores in the L2-L2 than the L2-L1 conditions were to some extent due to spelling errors. We have used the most lenient rating of the free recall protocols for language mistakes as we possibly could, because our goal was to get the best understanding of the memory representation rather than the quality of written output. If language mistakes had been punished, a larger L2 production cost would have been found in our study as well.

An interesting case in this respect was observed in the text about the Sun. As we described in the method section, three ideas contained the word billion, which is not the translation equivalent of the Dutch biljoen but of miljard. Many participants translated this to the false friend biljoen in their Dutch recall protocol, hence overestimating the Sun's age by a factor of 1000. We turned a blind eye to this mistake, because false friends
between languages are infrequent and it was clear the participants had retained the information, but on a real exam such a mistake would be punished (or more probably prevented by the lecturer pointing out several times in lectures the problem with the translation of the word billion). More importantly, depending on how the error is treated, the scores in the L2-to-L1 conditions could have been some 10% lower than the ones shown in Table 4, due to one specific false friend (occurring three times). At the same time, the word billion beautifully illustrates a real risk of using a non-native language for instruction: when students come across false friends, they may not be aware of their deficient knowledge of vocabulary and may not discover this discrepancy unless their attention is directed towards it by teachers or handbooks. Such mistakes will lead to lower scores on actual exams, in which the grading will be less lenient towards language mistakes.

*Billion* is in interesting example for another reason, because it can be considered an example of sub-technical or academic vocabulary (Nation, 2013a, 2013b): words you need to understand academic and domain-specific texts, but that are not domain-specific themselves (though *billion* happens to be a really frequent word compared to many other sub-technical or academic terms). Gablasova (2015) argues that academic vocabulary is problematic in non-native academic instruction, since it leads to ‘gaps’ in newly acquired knowledge of technical (or domain-specific) words. In her study (Gablasova, 2014, 2015), participants did acquire such terms that were defined in a text in L2, but less so than in L1 (and in less depth; note that this was tested half in L1 and half in L2 and that the author does not include the language of recall in the analysis).

Knowledge gaps are indeed the most likely explanation for the results we obtained. Because of knowledge gaps, the mental model of the text is weaker. This means the encoding or retrieval of the information is problematic. Since little or no evidence of language-dependency is reflected in the pattern of results, we can assume a language non-selective framework has to account for this, and indeed, two adequate frameworks are available. Firstly, the resource hypothesis states that processing L2 materials is more demanding and thus leaves less working memory capacity for higher-order processes (Sandoval et al., 2010). In the context of learning from text, higher-order processes are required for integrating the knowledge into prior knowledge, making inferences, and probably also actively encoding it in memory. Since a time limit is used here, readers
cannot compensate for this resource limitation in L2. Secondly, the weaker-links hypothesis assumes that linguistic representations of L2 words are weaker and their semantic representations less detailed, since familiarity with them is lower (comparable to low-frequent words in L1; Finkbeiner, 2002; Gollan, Montoya, Cera, & Sandoval, 2008). When readers process a text, they continuously build and update a mental model of the text (see the landscape model by van den Broek, Young, Tzeng, & Linderholm, 1999). Words do not only activate their own semantics but also other words, resulting in cohort activation. If semantic representations are less detailed and there are fewer connections between words, the cohort activation is less strong. Moreover, the occurrence of unknown words will not lead to any activation at all, which are similar to the ‘gaps’ Gablasova referred to (2015). This process would predict weaker recall, especially because active retrieval is more demanding than recognition and will suffer more from weaker representations in the model (representations of marginal knowledge, see Berger, Hall, & Bahrick, 1999; Cantor et al., 2014). According to some authors, the weaker-links hypothesis predicts better recognition memory in L2 since these memory traces are more unique (Francis & Gutiérrez, 2012). This assumption is not supported by the evidence in the current paper, but it might be limited to (episodic) memory for word lists.

In general, we saw no advantage for language congruency in recall in our studies. The scores were not better when texts and tests were in the same language than when they were in different languages, except perhaps in a recognition test. This is in line with the idea that text meanings, as soon as access to meaning is relevant or primed (by a word list, or by the task requirements), are not stored in a language-dependent way, but as language-independent, abstract propositions.

Though the present results are consistent with our previous research (in particular that with the same text materials), there are some limitations that require carefulness in interpretation. One is that the recognition tests were administered after the free recall tests, and though no inflation of the scores was observed, this may have resulted in the use of knowledge processed at a deeper level as part of the free recall test. If this is true, it might be possible to find a consistent and larger language-congruency advantage if the information can only be processed superficially (given that a language-congruency advantage is mainly reported for episodic memory; Francis & Gutiérrez, 2012). Another
limitation of the present study was that vocabulary support was given, which could also change the way the L2 propositions were processed (less superficially since the meaning of the words was supported and perhaps even primed and, thus, less prone to language-depending). Finally, there are indications that the results may be partially text-dependent (Chapter 4). Working with carefully constructed and well-validated stimulus materials makes the results of various studies more consistent, but risks to be biased by text-specific features (as we have seen for the billion example). So, access to a wider range of stimulus materials would certainly be a good addition to the present research line.

To conclude, the evidence reported in this study points toward the resource hypothesis or the weaker-links hypothesis of language non-selective access, and does not provide any clear-cut evidence for language-dependent semantic memory. The practical implication is that offering difficult, expository information to students entails a risk which is not entirely solved by testing them in their native language. It seems that the challenge is in understanding the materials to a deeper level in the first place, at least under time pressure and in this population. In addition, support for academic vocabulary and attention for false friends is necessary to increase success in (higher educational) foreign language learning situations. Future research could exclude the option that recognition memory is language-dependent by using recognition tests as the first or only measure for memory. Other studies could further explore the difficulties in encoding, retrieval and production in L2, taking factors such as difficulty, time, and vocabulary knowledge into account.
References


APPENDIX

Table 5. Reliability and correlations of the proficiency and WM measures in Experiment 1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng. LexTALE</td>
<td>0.79</td>
<td>0.57</td>
<td>0.71</td>
<td>0.47</td>
<td>0.11</td>
<td>0.19</td>
</tr>
<tr>
<td>Eng. spelling</td>
<td>0.70</td>
<td>0.83</td>
<td>0.54</td>
<td>0.49</td>
<td>0.46</td>
<td>0.47</td>
</tr>
<tr>
<td>Oxford QPT</td>
<td>0.93</td>
<td>0.69</td>
<td>0.74</td>
<td>0.5</td>
<td>0.04</td>
<td>0.42</td>
</tr>
<tr>
<td>Dutch voc.</td>
<td>0.57</td>
<td>0.58</td>
<td>0.62</td>
<td>0.87</td>
<td>0.44</td>
<td>0.17</td>
</tr>
<tr>
<td>Dutch spelling</td>
<td>0.16</td>
<td>0.64</td>
<td>0.06</td>
<td>0.59</td>
<td>0.63</td>
<td>0.21</td>
</tr>
<tr>
<td>Operation Span</td>
<td>0.24</td>
<td>0.57</td>
<td>0.54</td>
<td>0.20</td>
<td>0.29</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Note: On the diagonal (in italic) is the Cronbach’s alpha of each test. All numbers above that are original Pearson correlations. The numbers below the diagonal are the correlations corrected for reliability \((r_{xy}/\sqrt{(r_{xx}.r_{yy})})\). Missing data points were omitted by pairwise deletion.

Table 6. Reliability and correlations of the proficiency and WM measures in Experiment 2.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng. LexTALE</td>
<td>0.76</td>
<td>0.55</td>
<td>0.67</td>
<td>0.47</td>
<td>0.17</td>
<td>0.2</td>
</tr>
<tr>
<td>Eng. spelling</td>
<td>0.66</td>
<td>0.92</td>
<td>0.33</td>
<td>0.49</td>
<td>0.63</td>
<td>0.25</td>
</tr>
<tr>
<td>Oxford QPT</td>
<td>0.88</td>
<td>0.39</td>
<td>0.76</td>
<td>0.42</td>
<td>-0.09</td>
<td>0.27</td>
</tr>
<tr>
<td>Dutch voc.</td>
<td>0.61</td>
<td>0.57</td>
<td>0.54</td>
<td>0.79</td>
<td>0.32</td>
<td>0.19</td>
</tr>
<tr>
<td>Dutch spelling</td>
<td>0.22</td>
<td>0.74</td>
<td>-0.12</td>
<td>0.41</td>
<td>0.79</td>
<td>0.2</td>
</tr>
<tr>
<td>Op. Span</td>
<td>0.26</td>
<td>0.29</td>
<td>0.35</td>
<td>0.24</td>
<td>0.25</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Note: On the diagonal (in italic) is the cronbach’s alpha of each test. All numbers above that are original Pearson correlations. The numbers below the diagonal are the correlations corrected for reliability \((r_{xy}/\sqrt{(r_{xx}.r_{yy})})\). Missing data points were omitted by pairwise deletion.
6

EXPLORATORY ANALYSES:

ATTITUDES TOWARDS

ENGLISH AS A MEDIUM OF INSTRUCTION

AND FACTORS PREDICTING MEMORY FOR TEXTS
ABSTRACT

The aim of this chapter is to explore data gathered in this dissertation related to two questions. These questions were not central to the previous empirical chapters, but can improve our understanding of reading or studying texts in a second language (in this student population). The first question is about students’ motivation regarding the tests, academic reading in L1 and L2, their opinion about the use of English as a foreign language at university, and the strategies they use when encountering unknown vocabulary. Descriptives are reported for nearly 500 participants. Though reading motivation is generally higher in L1 than in L2, students are mostly positive about the use of EMI. The second question is which factors predict memory for texts in a second language. The multitude of proficiency (and other) measures that were collected in Chapter 2 are included in linear regression models for memory test scores in L1 and L2 (separately) and selected via a stepwise procedure. In addition, the correlation matrices are given for these measures per memory test condition. Though the models explain little variance, some interesting differences are discussed. The fact that working memory is a significant predictor for L2 tests and not for L1 tests confirms that working memory capacity is taxed more in L2, which is in line with the resource hypothesis.
1. What are students’ attitudes towards EMI in higher education?

The decision to use a foreign language for instruction affects students in the first place. Therefore, an important question apart from memory is what the students’ opinion is about this choice. Do they consider English as a useful asset for their further career or as a stumbling block for their studies? In addition, motivation is an important predictor for reading comprehension (Liebfreund, 2015; Schwabe, McElvany, & Trendtel, 2014). If motivation for reading in a non-native language differs from that in a native language, it will thus affect performance.

In all experimental set-ups in this dissertation, a questionnaire was included to inform about participants’ attitudes and motivation towards the test, towards reading in English and Dutch, and their experience of the text. To get an idea of how (mainly first-year) students think about EMI, an overview of these measures will be given.

1.1 Overview of attitude and motivation ratings

In Chapter 2, the questionnaire was limited to some single-item questions (see Appendix D.1). In Chapters 3 and 5, we elaborated the questionnaire and used sum scores (based on 3 or 4 questions) to achieve a more robust estimation (see Appendix D.2). The questions tapped into test motivation, intrinsic reading motivation, attitudes toward academic reading, extrinsic motivation towards academic reading, and opinions about EMI. Answers were given by selecting a number on a 7-point Likert scale ranging from “not at all”, “not difficult” … or a similar expression depending on the question, to “very much so”, “very difficult”, … Table 1 provides the means, standard deviations, and modes throughout the chapters. Since Likert-

---

1 These questions were inspired by a Questionnaire for academic reading and listening in a foreign language developed by Sari Verachtert, Elise De Mets, Ellen De Bruyne and Martin Valcke (see...). We are grateful to these authors for sharing their work with us (Demets, De Bruyne, & Valcke, 2016; Verachtert, De Bruyne, & Valcke, 2016).
scales (and individual Likert-scaled items) are not continuous in nature, we believe the mode provides a better idea of what most students actually responded, but for the sake of interpretation, we also report means and standard deviations. For the sum scores (Chapter 3 and 5), we have reported the total mode of all questions that were included in that sum score. For the individual Likert-scaled items (from Chapter 2), frequency histograms are included at the end of this chapter (Appendix 1). This can help avoid a mean score to be taken for evidence of neutral motivation, while it might be the consequence of many subjects selecting the middle point of the scale as an ‘escape option’. For the full Likert-scales (sum scores Chapter 3 and 5), histograms are also included, but in this case, they denote all sum scores around an integer (1-1.5, 1.5-2.5, 2.5-3.5 …), instead of actual responses within a category.

Though the questionnaire was limited, the large number of participants allows us to compare the motivation for reading (mostly academic texts) in Dutch and in English. It can be derived from Table 1 that, in most cases, motivation is higher in Dutch, and this was confirmed by Wilcoxon signed rank tests in Chapter 2 and 3. Note that these questions were asked AFTER testing. Hence, the experience that the task was more difficult in L2 might be reflected in the reduced motivation ratings as well. In addition, mean motivation (and mode) are slightly positive in English, which is good news for the use of EMI (motivation partially predicts performance). The fact that subjects were around a score of 5 shows they were motivated enough to participate in the experiment. Their motivation is probably higher on real-life tests, but still, this result is reassuring for the generalisability of our findings. The measure “performance” refers to a question in which subjects were asked to estimate how well they performed in comparison to their peers. On average, subjects thought to have performed in a comparable way.

The reason no IRT-analyses were done here, is that the number of items and subjects was rather limited to use advanced IRT-techniques. In the motivational measures for Chapter 3 and 5, reliability analyses were done on the subsets of questions created to measure a certain variable, such as intrinsic motivation. In addition, a principal component analysis has been done to check whether certain items loaded particularly strong on a certain component. Since this was not the case, because the first component was the only component which explained a substantial amount of variance, we kept to the initial categories of questions tapping into intrinsic motivation and attitudes, and included all questions (some tapping into extrinsic motivation, which seemed to be less effective) in a total motivation score. These questions might be improved with further validation. Still, they provide a nice overview of impressions of the subjects, and clear differences between languages show their effectiveness. The same items were used for Dutch and English reading motivation sum scores.
Furthermore, the questionnaire explicitly asked about opinions of the students towards EMI. Three questions asked whether students find the use of English for academic texts in their studies sensible (justified), a surplus, and inevitable. One question was formulated in a negative sense (hence, scores were mirrored), asking whether they found EMI for texts impermissible. As the results on these summed measures show, our participants were rather positive towards the use of English for academic texts, with 7 as the most frequent category. In addition, it is clear that the results are very steady across studies.

Table 1. Overview of the means, standard deviations and modes for single-item Likert scales and sum scores for motivation throughout this dissertation.

<table>
<thead>
<tr>
<th>Self-ratings</th>
<th>Chapter 2 (N = 195)</th>
<th>Chapter 3 (N = 155)</th>
<th>Chapter 5³ (N = 142)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General motivation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test importance (7)</td>
<td>5.09 (1.03)</td>
<td>5</td>
<td>4.81 (1.27)</td>
</tr>
<tr>
<td>Performance vs. peers (7)</td>
<td>4.22 (0.82)</td>
<td>4</td>
<td>3.90 (0.75)</td>
</tr>
<tr>
<td>Dutch reading</td>
<td>5.18 (1.41)</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Attitude (7)*</td>
<td>4.67 (1.11)</td>
<td>5</td>
<td>5.19 (1.05)</td>
</tr>
<tr>
<td>Intr. acad. motivation (7)*</td>
<td>4.89 (0.97)</td>
<td>5</td>
<td>4.99 (0.96)</td>
</tr>
<tr>
<td>Total acad. motivation (7)*</td>
<td>5.07 (0.81)</td>
<td>5</td>
<td>5.12 (0.82)</td>
</tr>
<tr>
<td><strong>English reading</strong></td>
<td>4.51 (1.47)</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Attitude (7)*</td>
<td>5.65 (0.95)</td>
<td>5</td>
<td>5.11 (1.13)</td>
</tr>
<tr>
<td>Intr. acad. motivation (7)*</td>
<td>4.34 (1.07)</td>
<td>5</td>
<td>4.60 (1.03)</td>
</tr>
<tr>
<td>Total acad. motivation (7)*</td>
<td>4.46 (0.98)</td>
<td>5</td>
<td>4.70 (0.90)</td>
</tr>
<tr>
<td>Opinion use of EMI (7)*</td>
<td>5.36 (1.16)</td>
<td>7</td>
<td>5.98 (0.91)</td>
</tr>
</tbody>
</table>

³ In chapter 4, 26 subjects were removed because of an error, and 26 new participants were tested. Since this error did not have anything to do with the current question, we do include their motivational data here.
1.2 Coping with unknown vocabulary

Apart from this recurring motivation questionnaire, we were curious to know how students cope with unknown vocabulary they encounter in an English text. This was added to the questionnaire for experiment 2 of Chapter 5. The following answering options were given (in Dutch, but we translated them here):

- I look up the translation in a translation dictionary (in print or online)
- I look up the (English) explanation in an English dictionary (in print or online)
- I ask someone
- I type the word in Google or another search engine
- I derive the meaning from the context
- I ignore the word
- I write it down to look it up later
- Other (write down in the next question)

Subjects were asked to order these strategies according to what they use the most, selecting at least one strategy. Table 2 summarizes the answers by displaying the counts of every strategy in every choice rank (1 to 8).

Table 2. Counts of the rankings subjects provided for strategies they use when encountering unknown vocabulary.

<table>
<thead>
<tr>
<th>Strategies</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>All ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation dictionary</td>
<td>9</td>
<td>18</td>
<td>23</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td>Explanation (standard dictionary)</td>
<td>3</td>
<td>12</td>
<td>16</td>
<td>16</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>0</td>
<td>67</td>
</tr>
<tr>
<td>Ask someone</td>
<td>3</td>
<td>8</td>
<td>8</td>
<td>12</td>
<td>17</td>
<td>9</td>
<td>10</td>
<td>0</td>
<td>67</td>
</tr>
<tr>
<td>Google/search engine</td>
<td>15</td>
<td></td>
<td>30</td>
<td>20</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>79</td>
</tr>
<tr>
<td>Derive from context</td>
<td>54</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>84</td>
</tr>
<tr>
<td>Ignore</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>8</td>
<td>17</td>
<td>14</td>
<td>10</td>
<td>1</td>
<td>61</td>
</tr>
<tr>
<td>Procrastinate (look up later)</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>18</td>
<td>16</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>32</td>
<td>34</td>
</tr>
<tr>
<td>All</td>
<td>84</td>
<td>84</td>
<td>80</td>
<td>72</td>
<td>64</td>
<td>49</td>
<td>46</td>
<td>34</td>
<td>513</td>
</tr>
</tbody>
</table>
Interestingly, the strategy that was chosen most on rank 1 is to derive the word from its context (54 out 84 subjects). After that, translation dictionaries and search engines are quite popular. Though they do not appear in the first ranks, in total, asking someone or looking it up in a traditional dictionary is also quite popular. In the last ranks, ignoring and procrastinating show up, but it is possible that subjects believed they had to rank all optional strategies, even when they never used them. Therefore, the top 3 strategies are more relevant. In general, it is clear that many different strategies are used. The results show that readers make an effort to understand a word, but if they believe they can derive it from the context, they will not check their interpretation. This might lead to misunderstandings, and as a consequence, to wrong elements in the mental model. This confirms our advice expressed in Chapter 5: vocabulary support is important, especially for false friends and subtechnical vocabulary. In this respect, the findings of Barry and Lazarte (1998) are interesting. They investigated how factors such as prior knowledge and syntactic complexity affect inference generation in a non-native recall task. For this purpose, they annotated within-text inferences, elaborative inferences and incorrect inferences. Incorrect inferences either contradict the actual text information or contain wrong interpretations of the surface code. From the examples given by Barry and Lazarte, it becomes clear that such incorrect inferences are often caused by wrong vocabulary understanding. One such example they frequently encountered was the misinterpretation of the Spanish guarder (to guard or to store) as a noun instead of a verb, resulting in a proposition about guards near a building, while the original proposition explained that grain was stored in a building. We can assume that such misunderstandings occur more when readers derive a word from its context.

Only 6 subjects referred to another strategy. One subject uses google translate to translate a part of the text and another subject looks for information on the internet – both of these are similar to our categories (translation and search engine). The only other strategies mentioned were to derive the word from another language (3 subjects) or to replace the word by another word (which is a rather vague answer).

To sum up: reading motivation seems to be a bit higher in Dutch, at least after the experimental tasks. Furthermore, students were not very but sufficiently motivated for our empirical investigation. Their (average) opinion towards EMI is positive. Lastly, they try to derive the meaning of unknown words from their context more than looking them up.
2. Which factors predict memory for texts?

The main factors that predict reading comprehension were mentioned in the introduction. It was also discussed that language-related factors such as decoding ability are more important predictors in a non-native language. Throughout the current dissertation, several tests measuring L1 proficiency, L2 proficiency, and working memory were administered. The main reason was to compare groups and to estimate the general level of the entire participant pool, but these measures can also be used to explore which are the most predictive of memory for text in certain situations.

In Chapter 2, a multitude of proficiency measures was administered, which allows us to run exploratory analyses using linear regression on the memory scores. Memory tests were administered in language groups with test type as a repeated measure. Therefore, we will report on linear regressions per language and per test type. We will report full models, since all factors are expected to influence reading comprehension and memory, and partial models, by stepwise selection of factors. Obviously, factors such as receptive vocabulary knowledge (LexTALE) and spelling are correlated (see correlation tables in every chapter). Hence, stepwise selection can rule out factors that are redundant. The full model was always compared to the result of the stepwise selection procedure by anova comparison, but this never reached significance. Adjusted r-squared (r²) is reported for every model, which indicates how much variance is explained by the model.

On the following page, the full model for every test type is given, followed by the selected model. Factors that reached significance are underlined.
**Full: Free recall L1 ~** Dutch LexTALE + Dutch vocabulary (MC) + Dutch spelling + Dutch reading motivation + Author test + Working memory + IQ

(Note: Dutch spelling: p = .07) $r^2 = .09$

**Stepwise: Free recall L1 ~** Dutch vocabulary (MC) + Dutch spelling + IQ $r^2 = .11$

**Full: Free recall L2 ~** English LexTALE + Oxford QPT + MINT + English spelling + English reading motivation + Author test + Working memory + IQ

$\text{r}^2 = .27$

**Stepwise: Free recall L2 ~** Oxford QPT + MINT + English spelling + English reading motivation + Author test + Working memory

$r^2 = .29$

**Full: True/false L1 ~** Dutch LexTALE + Dutch vocabulary (MC) + Dutch spelling + Dutch reading motivation + Author test + Working memory + IQ

$\text{r}^2 = .06$

**Stepwise: True/false L1 ~** Dutch LexTALE + Dutch vocabulary (MC) $\text{r}^2 = .07$

**Full: True/false L2 ~** English LexTALE + Oxford QPT + MINT + English spelling + English reading motivation + Author test + Working memory + IQ

(Note: MINT: p = .06) $\text{r}^2 = .18$

**Stepwise: True/false L2 ~** English LexTALE + MINT + Working memory

(Note: other variables at p = .08) $\text{r}^2 = .19$
Though the stepwise selected models do not perform significantly better than the full models, it is interesting to see which factors come out. Firstly, the MINT picture naming task is a significant (or near-significant) predictor in all L2 tasks. It might rule out the effect of other L2 proficiency measures. In L1 recall, Dutch spelling is the only significant predictor, while in L2 recall it is the MINT task in combination with working memory. The fact that working memory is only significant in the L2 recall task, indicates that working memory is taxed more in L2. This is in line with the resource hypothesis (Sandoval, Gollan, Ferreira, & Salmon, 2010). Participants who have more working memory capacity could perhaps deal with the demanding L2 processes better, while a higher MINT score points towards stronger active language knowledge (especially since this was administered in groups, resulting in time pressure for this task). In L1, while Dutch spelling predicts recall best, it is Dutch receptive vocabulary knowledge which predicts the recognition scores. This reflects the type of task: a productive measure versus a receptive measure. In general, test scores in L2 are predicted far better than in L1. Not only did we have more proficiency measures at our disposal in L2 than in L1; since L2 is a weaker language, proficiency measures will have a larger effect (lower scores and more variance).

A possible issue with this approach is that many of the variables are correlated. Still, to know which element is most important, the regressions provide us with most information. To have an idea of multicollinearity, correlation matrices are displayed in Tables 3 to 6 (at the end of this chapter, Appendix 2), including all independent variables and the dependent variable depending on the memory test group (e.g. L1 recall). Since 15 variables are included, a Bonferroni correction for multiple testing was applied, resulting in an alpha level of 0.00042. Significant p-values (below the diagonal) are highlighted. The correlation matrices show that English proficiency measures significantly correlate to each other, sometimes including English reading motivation.
References


Appendix 1

Histograms Chapter 2

Frequency of responses on single-item Likert scales

Test importance

Performance vs. peers

Dutch reading motivation

English reading motivation
**Histograms Chapter 3**

**Frequency of responses on single-item Likert scales**

- **Test importance**
- **Performance vs. peers**

**Distribution of sum scores based on Likert scales**

- **Dutch RM: attitudes**
- **English RM: attitudes**

- **Dutch RM: intrinsic motivation**
- **English RM: intrinsic motivation**
EXPLORATORY ANALYSES

Dutch RM: total acad. motivation

English RM: total acad. motivation

Opinions on EMI
CHAPTER 6

Histograms Chapter 5

**Frequency of responses on single-item Likert scales**

- **Test importance**
  - Frequency distribution
  - Bars for each 1-7 rating

- **Performance vs. peers**
  - Frequency distribution
  - Bars for each 1-7 rating

**Distribution of sum scores based on Likert scales**

- **Dutch RM: attitudes**
  - Frequency distribution
  - Bars for each 1-7 rating

- **English RM: attitudes**
  - Frequency distribution
  - Bars for each 1-7 rating

- **Dutch RM: intrinsic motivation**
  - Frequency distribution
  - Bars for each 1-7 rating

- **English RM: intrinsic motivation**
  - Frequency distribution
  - Bars for each 1-7 rating
Appendix 2
<table>
<thead>
<tr>
<th>L1 True/False</th>
<th>Dutch LexTALE</th>
<th>Dutch MC vocabulary</th>
<th>Dutch spelling</th>
<th>Dutch RM</th>
<th>English LexTALE</th>
<th>Oxford QPT</th>
<th>MINT</th>
<th>English Spelling</th>
<th>English RM</th>
<th>Author test</th>
<th>WM</th>
<th>IQ</th>
<th>Test importance</th>
<th>Selfrating</th>
<th>Memory score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch LexTALE</td>
<td>NA</td>
<td>0.255</td>
<td>0.155</td>
<td>0.035</td>
<td>0.288</td>
<td>0.235</td>
<td>0.194</td>
<td>0.217</td>
<td>0.041</td>
<td>0.183</td>
<td>0.055</td>
<td>-0.005</td>
<td>-0.095</td>
<td>-0.185</td>
<td>0.206</td>
</tr>
<tr>
<td>Dutch MC vocabulary</td>
<td>0.011</td>
<td>NA</td>
<td>0.263</td>
<td>-0.028</td>
<td>0.253</td>
<td>0.381</td>
<td>0.340</td>
<td>0.340</td>
<td>0.106</td>
<td>0.369</td>
<td>0.151</td>
<td>0.223</td>
<td>0.042</td>
<td>0.039</td>
<td>0.256</td>
</tr>
<tr>
<td>Dutch spelling</td>
<td>0.127</td>
<td>0.009</td>
<td>NA</td>
<td>-0.028</td>
<td>0.286</td>
<td>0.368</td>
<td>0.252</td>
<td>0.618</td>
<td>0.217</td>
<td>0.270</td>
<td>0.204</td>
<td>0.217</td>
<td>0.233</td>
<td>0.014</td>
<td>0.006</td>
</tr>
<tr>
<td>Dutch RM</td>
<td>0.729</td>
<td>0.784</td>
<td>0.206</td>
<td>NA</td>
<td>-0.117</td>
<td>0.000</td>
<td>-0.092</td>
<td>-0.052</td>
<td>0.146</td>
<td>0.059</td>
<td>-0.161</td>
<td>-0.181</td>
<td>0.151</td>
<td>0.190</td>
<td>-0.130</td>
</tr>
<tr>
<td>English LexTALE</td>
<td>0.004</td>
<td>0.012</td>
<td>0.004</td>
<td>0.250</td>
<td>NA</td>
<td>0.678</td>
<td>0.722</td>
<td>0.574</td>
<td>0.399</td>
<td>0.243</td>
<td>0.269</td>
<td>0.198</td>
<td>-0.076</td>
<td>-0.050</td>
<td>0.227</td>
</tr>
<tr>
<td>Oxford QPT</td>
<td>0.020</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>1.000</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.752</td>
<td>0.580</td>
<td>0.618</td>
<td>0.230</td>
<td>0.273</td>
<td>0.194</td>
<td>-0.052</td>
<td>-0.027</td>
<td>0.240</td>
</tr>
<tr>
<td>MINT</td>
<td>0.056</td>
<td>0.001</td>
<td>0.012</td>
<td>0.368</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.543</td>
<td>0.553</td>
<td>0.151</td>
<td>0.165</td>
<td>0.150</td>
<td>-0.084</td>
<td>0.007</td>
<td>0.342</td>
</tr>
<tr>
<td>English Spelling</td>
<td>0.032</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>0.608</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.413</td>
<td>0.279</td>
<td>0.325</td>
<td>0.121</td>
<td>0.125</td>
<td>-0.010</td>
<td>0.107</td>
</tr>
<tr>
<td>English RM</td>
<td>0.691</td>
<td>0.298</td>
<td>0.032</td>
<td>0.150</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.131</td>
<td>0.056</td>
<td>-0.021</td>
<td>-0.008</td>
<td>0.125</td>
<td>-0.035</td>
<td></td>
</tr>
<tr>
<td>Author test</td>
<td>0.071</td>
<td>&lt;.00042</td>
<td>0.007</td>
<td>0.564</td>
<td>0.016</td>
<td>0.022</td>
<td>0.139</td>
<td>0.005</td>
<td>0.199</td>
<td>NA</td>
<td>0.162</td>
<td>0.138</td>
<td>0.061</td>
<td>-0.070</td>
<td>0.021</td>
</tr>
<tr>
<td>WM</td>
<td>0.589</td>
<td>0.139</td>
<td>0.044</td>
<td>0.114</td>
<td>0.007</td>
<td>0.007</td>
<td>0.105</td>
<td>0.001</td>
<td>0.581</td>
<td>0.111</td>
<td>NA</td>
<td>0.067</td>
<td>0.084</td>
<td>0.038</td>
<td>-0.049</td>
</tr>
<tr>
<td>IQ</td>
<td>0.962</td>
<td>0.027</td>
<td>0.031</td>
<td>0.074</td>
<td>0.051</td>
<td>0.056</td>
<td>0.142</td>
<td>0.236</td>
<td>0.836</td>
<td>0.174</td>
<td>0.514</td>
<td>NA</td>
<td>0.112</td>
<td>-0.013</td>
<td>0.096</td>
</tr>
<tr>
<td>Test importance</td>
<td>0.353</td>
<td>0.684</td>
<td>0.021</td>
<td>0.139</td>
<td>0.456</td>
<td>0.614</td>
<td>0.412</td>
<td>0.222</td>
<td>0.936</td>
<td>0.548</td>
<td>0.414</td>
<td>0.273</td>
<td>NA</td>
<td>0.181</td>
<td>0.008</td>
</tr>
<tr>
<td>Selfrating</td>
<td>0.068</td>
<td>0.703</td>
<td>0.893</td>
<td>0.060</td>
<td>0.626</td>
<td>0.794</td>
<td>0.943</td>
<td>0.923</td>
<td>0.222</td>
<td>0.493</td>
<td>0.709</td>
<td>0.900</td>
<td>0.075</td>
<td>NA</td>
<td>0.071</td>
</tr>
<tr>
<td>Memory score</td>
<td>0.042</td>
<td>0.011</td>
<td>0.956</td>
<td>0.202</td>
<td>0.024</td>
<td>0.017</td>
<td>0.001</td>
<td>0.294</td>
<td>0.733</td>
<td>0.834</td>
<td>0.634</td>
<td>0.345</td>
<td>0.934</td>
<td>0.485</td>
<td>NA</td>
</tr>
</tbody>
</table>

**Table 3.** Correlation matrix of covariates for true/false in L1. Pearson correlation is displayed above the diagonal, p-values are displayed below the diagonal. After bonferroni correction, the significance level is at 0.00042.
<table>
<thead>
<tr>
<th>L2 True/False</th>
<th>Dutch LexTALE</th>
<th>Dutch MC vocabulary</th>
<th>Dutch spelling</th>
<th>Dutch RM</th>
<th>English LexTALE</th>
<th>Oxford QPT</th>
<th>MINT</th>
<th>English Spelling</th>
<th>English RM</th>
<th>Author test</th>
<th>WM</th>
<th>IQ</th>
<th>Test importance</th>
<th>Selfrating</th>
<th>Memory score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch LexTALE</td>
<td>NA</td>
<td>0.019</td>
<td>0.243</td>
<td>0.051</td>
<td>0.395</td>
<td>0.206</td>
<td>0.259</td>
<td>0.252</td>
<td>-0.059</td>
<td>0.118</td>
<td>0.153</td>
<td>0.104</td>
<td>-0.060</td>
<td>0.042</td>
<td>0.090</td>
</tr>
<tr>
<td>Dutch MC vocabulary</td>
<td>0.850</td>
<td>NA</td>
<td>0.182</td>
<td>-0.014</td>
<td>0.279</td>
<td>0.390</td>
<td>0.275</td>
<td>0.291</td>
<td>0.120</td>
<td>0.309</td>
<td>-0.076</td>
<td>0.159</td>
<td>0.055</td>
<td>-0.024</td>
<td>0.191</td>
</tr>
<tr>
<td>Dutch spelling</td>
<td>0.016</td>
<td>0.075</td>
<td>NA</td>
<td>-0.113</td>
<td>0.307</td>
<td>0.324</td>
<td>0.302</td>
<td>0.742</td>
<td>0.146</td>
<td>0.307</td>
<td>0.074</td>
<td>0.110</td>
<td>0.056</td>
<td>0.174</td>
<td>0.162</td>
</tr>
<tr>
<td>Dutch RM</td>
<td>0.618</td>
<td>0.895</td>
<td>0.272</td>
<td>NA</td>
<td>-0.204</td>
<td>-0.163</td>
<td>-0.205</td>
<td>-0.226</td>
<td>0.047</td>
<td>-0.070</td>
<td>-0.160</td>
<td>-0.175</td>
<td>-0.005</td>
<td>-0.128</td>
<td>-0.064</td>
</tr>
<tr>
<td>English LexTALE</td>
<td>&lt;.00042</td>
<td>0.006</td>
<td>0.002</td>
<td>0.045</td>
<td>NA</td>
<td>0.497</td>
<td>0.506</td>
<td>0.522</td>
<td>0.291</td>
<td>0.300</td>
<td>0.086</td>
<td>0.247</td>
<td>-0.051</td>
<td>0.188</td>
<td>0.350</td>
</tr>
<tr>
<td>Oxford QPT</td>
<td>0.043</td>
<td>&lt;.00042</td>
<td>0.001</td>
<td>0.111</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.539</td>
<td>0.540</td>
<td>0.404</td>
<td>0.257</td>
<td>-0.053</td>
<td>0.129</td>
<td>0.066</td>
<td>0.238</td>
<td>0.292</td>
</tr>
<tr>
<td>MINT</td>
<td>0.010</td>
<td>0.007</td>
<td>0.003</td>
<td>0.043</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.577</td>
<td>0.412</td>
<td>0.202</td>
<td>-0.168</td>
<td>0.221</td>
<td>0.035</td>
<td>0.311</td>
<td>0.421</td>
</tr>
<tr>
<td>English Spelling</td>
<td>0.013</td>
<td>0.004</td>
<td>&lt;.00042</td>
<td>0.026</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.265</td>
<td>0.279</td>
<td>-0.040</td>
<td>0.162</td>
<td>-0.003</td>
<td>0.205</td>
<td>0.355</td>
</tr>
<tr>
<td>English RM</td>
<td>0.569</td>
<td>0.241</td>
<td>0.154</td>
<td>0.651</td>
<td>0.004</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>0.009</td>
<td>NA</td>
<td>0.048</td>
<td>-0.173</td>
<td>-0.061</td>
<td>0.176</td>
<td>0.295</td>
<td>0.232</td>
</tr>
<tr>
<td>Author test</td>
<td>0.251</td>
<td>0.002</td>
<td>0.002</td>
<td>0.047</td>
<td>0.003</td>
<td>0.011</td>
<td>0.047</td>
<td>0.006</td>
<td>0.639</td>
<td>NA</td>
<td>0.155</td>
<td>0.039</td>
<td>-0.091</td>
<td>-0.064</td>
<td>0.130</td>
</tr>
<tr>
<td>WM</td>
<td>0.138</td>
<td>0.463</td>
<td>0.472</td>
<td>0.119</td>
<td>0.406</td>
<td>0.609</td>
<td>0.101</td>
<td>0.699</td>
<td>0.091</td>
<td>0.131</td>
<td>NA</td>
<td>0.288</td>
<td>-0.001</td>
<td>0.004</td>
<td>0.091</td>
</tr>
<tr>
<td>IQ</td>
<td>0.311</td>
<td>0.121</td>
<td>0.286</td>
<td>0.087</td>
<td>0.015</td>
<td>0.207</td>
<td>0.030</td>
<td>0.112</td>
<td>0.555</td>
<td>0.703</td>
<td>0.004</td>
<td>NA</td>
<td>0.112</td>
<td>0.039</td>
<td>0.265</td>
</tr>
<tr>
<td>Test importance</td>
<td>0.559</td>
<td>0.593</td>
<td>0.584</td>
<td>0.964</td>
<td>0.622</td>
<td>0.519</td>
<td>0.735</td>
<td>0.975</td>
<td>0.085</td>
<td>0.373</td>
<td>0.994</td>
<td>0.273</td>
<td>NA</td>
<td>0.442</td>
<td>0.191</td>
</tr>
<tr>
<td>Selfrating</td>
<td>0.684</td>
<td>0.815</td>
<td>0.088</td>
<td>0.210</td>
<td>0.065</td>
<td>0.019</td>
<td>0.002</td>
<td>0.044</td>
<td>0.003</td>
<td>0.531</td>
<td>0.967</td>
<td>0.705</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.338</td>
</tr>
<tr>
<td>Memory score</td>
<td>0.382</td>
<td>0.060</td>
<td>0.113</td>
<td>0.531</td>
<td>0.000</td>
<td>0.004</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>0.022</td>
<td>0.206</td>
<td>0.378</td>
<td>0.009</td>
<td>0.061</td>
<td>0.001</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 4. Correlation matrix of covariates for true/false test in L2. Pearson correlation is displayed above the diagonal, p-values are displayed below the diagonal. After bonferroni correction, the significance level is at 0.00042.
<table>
<thead>
<tr>
<th>L1 Free recall</th>
<th>Dutch LexTALE</th>
<th>Dutch MC vocabulary</th>
<th>Dutch spelling</th>
<th>Dutch RM</th>
<th>English LexTALE</th>
<th>Oxford QPT</th>
<th>MINT</th>
<th>English Spelling</th>
<th>English RM</th>
<th>Author test</th>
<th>WM</th>
<th>IQ</th>
<th>Test importance</th>
<th>Selfrating</th>
<th>Memory score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch LexTALE</td>
<td>NA 0.255</td>
<td>0.155</td>
<td>0.035</td>
<td>0.288</td>
<td>0.235</td>
<td>0.194</td>
<td>0.217</td>
<td>0.041</td>
<td>0.183</td>
<td>0.055</td>
<td>-0.005</td>
<td>-0.095</td>
<td>-0.185</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>Dutch MC vocabulary</td>
<td>0.011  NA 0.263</td>
<td>-0.028</td>
<td>0.253</td>
<td>0.381</td>
<td>0.340</td>
<td>0.106</td>
<td>0.369</td>
<td>0.151</td>
<td>0.223</td>
<td>0.042</td>
<td>0.039</td>
<td>0.250</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch spelling</td>
<td>0.127 0.009</td>
<td>NA 0.129</td>
<td>0.286</td>
<td>0.368</td>
<td>0.252</td>
<td>0.618</td>
<td>0.217</td>
<td>0.270</td>
<td>0.204</td>
<td>0.217</td>
<td>0.233</td>
<td>0.014</td>
<td>0.285</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch RM</td>
<td>0.729</td>
<td>0.784</td>
<td>0.206</td>
<td>NA</td>
<td>0.117</td>
<td>0.000</td>
<td>-0.092</td>
<td>-0.052</td>
<td>0.146</td>
<td>0.059</td>
<td>-0.161</td>
<td>-0.181</td>
<td>0.151</td>
<td>0.190</td>
<td>-0.025</td>
</tr>
<tr>
<td>English LexTALE</td>
<td>0.004</td>
<td>0.012</td>
<td>0.004</td>
<td>0.250</td>
<td>NA 2E-14</td>
<td>0.722</td>
<td>0.574</td>
<td>0.399</td>
<td>0.243</td>
<td>0.269</td>
<td>0.198</td>
<td>-0.076</td>
<td>-0.050</td>
<td>0.203</td>
<td></td>
</tr>
<tr>
<td>Oxford QPT</td>
<td>0.020</td>
<td>&lt;.00042</td>
<td>1.000</td>
<td>NA</td>
<td>0.752</td>
<td>0.580</td>
<td>0.618</td>
<td>0.230</td>
<td>0.273</td>
<td>0.194</td>
<td>-0.052</td>
<td>-0.027</td>
<td>0.305</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MINT</td>
<td>0.056</td>
<td>0.001</td>
<td>0.012</td>
<td>0.368</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.543</td>
<td>0.553</td>
<td>0.151</td>
<td>0.165</td>
<td>0.150</td>
<td>-0.084</td>
<td>0.007</td>
<td>0.139</td>
</tr>
<tr>
<td>English Spelling</td>
<td>0.032</td>
<td>0.001</td>
<td>&lt;.00042</td>
<td>0.608</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.413</td>
<td>0.279</td>
<td>0.325</td>
<td>0.121</td>
<td>0.125</td>
<td>-0.010</td>
<td>0.213</td>
</tr>
<tr>
<td>English RM</td>
<td>0.691</td>
<td>0.298</td>
<td>0.032</td>
<td>0.150</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>&lt;.00042</td>
<td>NA</td>
<td>0.131</td>
<td>0.056</td>
<td>-0.021</td>
<td>-0.008</td>
<td>0.125</td>
<td>0.079</td>
</tr>
<tr>
<td>Author test</td>
<td>0.071</td>
<td>&lt;.00042</td>
<td>0.007</td>
<td>0.564</td>
<td>0.016</td>
<td>0.022</td>
<td>0.139</td>
<td>0.005</td>
<td>0.199</td>
<td>NA</td>
<td>0.162</td>
<td>0.138</td>
<td>0.061</td>
<td>-0.070</td>
<td>0.217</td>
</tr>
<tr>
<td>WM</td>
<td>0.589</td>
<td>0.139</td>
<td>0.044</td>
<td>0.114</td>
<td>0.007</td>
<td>0.007</td>
<td>0.105</td>
<td>0.001</td>
<td>0.581</td>
<td>0.111</td>
<td>NA</td>
<td>0.067</td>
<td>0.084</td>
<td>0.038</td>
<td>0.105</td>
</tr>
<tr>
<td>IQ</td>
<td>0.962</td>
<td>0.027</td>
<td>0.031</td>
<td>0.074</td>
<td>0.051</td>
<td>0.056</td>
<td>0.142</td>
<td>0.236</td>
<td>0.836</td>
<td>0.174</td>
<td>0.514</td>
<td>NA</td>
<td>0.112</td>
<td>-0.013</td>
<td>0.248</td>
</tr>
<tr>
<td>Test importance</td>
<td>0.353</td>
<td>0.684</td>
<td>0.021</td>
<td>0.139</td>
<td>0.456</td>
<td>0.614</td>
<td>0.412</td>
<td>0.222</td>
<td>0.936</td>
<td>0.548</td>
<td>0.414</td>
<td>0.273</td>
<td>NA</td>
<td>0.181</td>
<td>-0.028</td>
</tr>
<tr>
<td>Selfrating</td>
<td>0.068</td>
<td>0.703</td>
<td>0.893</td>
<td>0.060</td>
<td>0.626</td>
<td>0.794</td>
<td>0.943</td>
<td>0.923</td>
<td>0.222</td>
<td>0.493</td>
<td>0.709</td>
<td>0.900</td>
<td>0.075</td>
<td>NA</td>
<td>0.178</td>
</tr>
<tr>
<td>Memory score</td>
<td>0.042</td>
<td>0.011</td>
<td>0.956</td>
<td>0.202</td>
<td>0.024</td>
<td>0.017</td>
<td>0.001</td>
<td>0.294</td>
<td>0.733</td>
<td>0.834</td>
<td>0.634</td>
<td>0.014</td>
<td>0.783</td>
<td>0.080</td>
<td>NA</td>
</tr>
</tbody>
</table>

**Table 5.** Correlation matrix of covariates for free recall in L1. Pearson correlation is displayed above the diagonal, p-values are displayed below the diagonal. After bonferroni correction, the significance level is at 0.00042.
<table>
<thead>
<tr>
<th>L1 Free recall</th>
<th>Dutch LexTALE</th>
<th>Dutch MC vocabulary</th>
<th>Dutch spelling</th>
<th>Dutch RM</th>
<th>English LexTALE</th>
<th>Oxford QPT</th>
<th>MINT</th>
<th>English Spelling</th>
<th>English RM</th>
<th>Author test</th>
<th>WM</th>
<th>IQ</th>
<th>Test importance</th>
<th>Selfrating</th>
<th>Memory score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch LexTALE</td>
<td>NA</td>
<td>0.019</td>
<td>0.243</td>
<td>0.051</td>
<td>0.395</td>
<td>0.206</td>
<td>0.259</td>
<td>0.252</td>
<td>-0.059</td>
<td>0.118</td>
<td>0.153</td>
<td>0.104</td>
<td>-0.060</td>
<td>0.042</td>
<td>0.226</td>
</tr>
<tr>
<td>Dutch MC vocabulary</td>
<td>0.850</td>
<td>NA</td>
<td>0.182</td>
<td>-0.014</td>
<td>0.279</td>
<td>0.390</td>
<td>0.275</td>
<td>0.291</td>
<td>0.120</td>
<td>0.309</td>
<td>0.076</td>
<td>0.159</td>
<td>0.055</td>
<td>-0.024</td>
<td>0.167</td>
</tr>
<tr>
<td>Dutch spelling</td>
<td>0.016</td>
<td>0.075</td>
<td>NA</td>
<td>-0.113</td>
<td>0.307</td>
<td>0.324</td>
<td>0.302</td>
<td>0.742</td>
<td>0.146</td>
<td>0.307</td>
<td>0.074</td>
<td>0.110</td>
<td>0.056</td>
<td>0.174</td>
<td>0.303</td>
</tr>
<tr>
<td>Dutch RM</td>
<td>0.618</td>
<td>0.895</td>
<td>0.272</td>
<td>NA</td>
<td>-0.204</td>
<td>-0.163</td>
<td>-0.205</td>
<td>-0.226</td>
<td>0.047</td>
<td>-0.070</td>
<td>-0.160</td>
<td>-0.175</td>
<td>-0.005</td>
<td>-0.128</td>
<td>-0.147</td>
</tr>
<tr>
<td>English LexTALE</td>
<td>&lt;0.00042</td>
<td>0.006</td>
<td>0.002</td>
<td>0.045</td>
<td>NA</td>
<td>0.497</td>
<td>0.506</td>
<td>0.522</td>
<td>0.291</td>
<td>0.300</td>
<td>-0.086</td>
<td>0.247</td>
<td>0.051</td>
<td>0.188</td>
<td>0.240</td>
</tr>
<tr>
<td>Oxford QPT</td>
<td>0.043</td>
<td>&lt;0.00042</td>
<td>0.001</td>
<td>0.111</td>
<td>&lt;0.00042</td>
<td>NA</td>
<td>0.539</td>
<td>0.540</td>
<td>0.404</td>
<td>0.257</td>
<td>-0.053</td>
<td>0.129</td>
<td>0.066</td>
<td>0.238</td>
<td>0.343</td>
</tr>
<tr>
<td>MINT</td>
<td>0.010</td>
<td>0.007</td>
<td>0.003</td>
<td>0.043</td>
<td>&lt;0.00042</td>
<td>&lt;0.00042</td>
<td>NA</td>
<td>0.577</td>
<td>0.412</td>
<td>0.202</td>
<td>-0.168</td>
<td>0.221</td>
<td>0.035</td>
<td>0.311</td>
<td>0.396</td>
</tr>
<tr>
<td>English Spelling</td>
<td>0.013</td>
<td>0.004</td>
<td>&lt;0.00042</td>
<td>0.026</td>
<td>&lt;0.00042</td>
<td>&lt;0.00042</td>
<td>&lt;0.00042</td>
<td>&lt;0.00042</td>
<td>NA</td>
<td>0.265</td>
<td>-0.040</td>
<td>0.162</td>
<td>-0.003</td>
<td>0.205</td>
<td>0.386</td>
</tr>
<tr>
<td>English RM</td>
<td>0.569</td>
<td>0.241</td>
<td>0.154</td>
<td>0.651</td>
<td>0.004</td>
<td>&lt;0.00042</td>
<td>&lt;0.00042</td>
<td>&lt;0.00042</td>
<td>NA</td>
<td>0.048</td>
<td>-0.173</td>
<td>-0.061</td>
<td>0.176</td>
<td>0.295</td>
<td>0.038</td>
</tr>
<tr>
<td>Author test</td>
<td>0.251</td>
<td>0.002</td>
<td>0.002</td>
<td>0.497</td>
<td>0.003</td>
<td>0.011</td>
<td>0.047</td>
<td>0.006</td>
<td>0.639</td>
<td>NA</td>
<td>0.155</td>
<td>0.039</td>
<td>-0.091</td>
<td>-0.064</td>
<td>0.017</td>
</tr>
<tr>
<td>WM</td>
<td>0.138</td>
<td>0.463</td>
<td>0.472</td>
<td>0.119</td>
<td>0.406</td>
<td>0.609</td>
<td>0.101</td>
<td>0.699</td>
<td>0.091</td>
<td>0.131</td>
<td>NA</td>
<td>0.288</td>
<td>-0.001</td>
<td>0.004</td>
<td>0.261</td>
</tr>
<tr>
<td>IQ</td>
<td>0.311</td>
<td>0.121</td>
<td>0.286</td>
<td>0.087</td>
<td>0.015</td>
<td>0.207</td>
<td>0.030</td>
<td>0.112</td>
<td>0.555</td>
<td>0.703</td>
<td>NA</td>
<td>0.112</td>
<td>0.039</td>
<td>0.240</td>
<td></td>
</tr>
<tr>
<td>Test importance</td>
<td>0.559</td>
<td>0.593</td>
<td>0.584</td>
<td>0.964</td>
<td>0.622</td>
<td>0.519</td>
<td>0.735</td>
<td>0.975</td>
<td>0.085</td>
<td>0.373</td>
<td>0.994</td>
<td>0.273</td>
<td>NA</td>
<td>0.442</td>
<td>-0.023</td>
</tr>
<tr>
<td>Selfrating</td>
<td>0.684</td>
<td>0.815</td>
<td>0.088</td>
<td>0.210</td>
<td>0.065</td>
<td>0.019</td>
<td>0.002</td>
<td>0.044</td>
<td>0.003</td>
<td>0.531</td>
<td>0.967</td>
<td>0.705</td>
<td>&lt;0.00042</td>
<td>NA</td>
<td>0.213</td>
</tr>
<tr>
<td>Memory score</td>
<td>0.029</td>
<td>0.107</td>
<td>0.003</td>
<td>0.158</td>
<td>0.020</td>
<td>0.001</td>
<td>&lt;0.00042</td>
<td>&lt;0.00042</td>
<td>0.717</td>
<td>0.869</td>
<td>0.011</td>
<td>0.020</td>
<td>0.829</td>
<td>0.039</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 6. Correlation matrix of covariates for free recall in L2. Pearson correlation is displayed above the diagonal, p-values are displayed below the diagonal. After bonferroni correction, the significance level is at 0.00042.
7

GENERAL DISCUSSION
The current dissertation investigated the presence, form and size of disadvantages in a non-native language for learning from text, in subjects at a Flemish university. Such disadvantages would have highly relevant implications for the use of English as a medium of instruction at higher education institutions. Furthermore, the investigation of memory for texts in a second language can inform us about the functioning of bilingual memory and the question of whether semantic memory is language-dependent or not. In this discussion, the findings of all empirical chapters will be recapitulated and juxtaposed with related literature. Due to the interdisciplinary character of this topic, some studies from different fields are relevant to our research question. Therefore, we will attempt to provide an exhaustive overview of studies comparing memory for texts in a native and non-native language. Finally, we will pinpoint some limitations of our own research and of the literature and suggest directions for further research.

1. Findings

First and foremost, the findings reported in this dissertation provide evidence for a disadvantage in L2. Memory for texts is impaired in a non-native language, at least for some tasks. At the same time, not all studies report a disadvantage. First, a summary of the findings in each chapter is provided. Next, we will report on factors that seem to be crucial to understand the inconsistencies in the literature and in this dissertation. An overview of effect sizes will be given. Then, the theoretical implications for several types of memory and frameworks of L2-processing will be summarised.

1.1 Summary of results

In Chapter 2, both recognition and recall of texts were tested in a native versus a non-native language. Two language groups received a short expository text and a
subsequent test in either Dutch or English. The test could be a true/false judgement test (recognition) or a free recall task, in which it is asked to write down everything you remember from a text. Results show no language difference on the true/false test, but a large L2 disadvantage on the free recall task (Cohen’s $d = .86$). In short: a large L2 recall cost was found, while recognition performance was similar in L1 and L2.

In Chapter 3, the forgetting curves were investigated for L1 and L2 recognition memory. Participants received an immediate and delayed test, based on the same text materials as in Chapter 2. The delayed test was administered either after a day, a week, or a month. The forgetting curves look very similar in L1 and L2 (with absolute scores slightly lower in L2) and no language effect was found in a mixed logistic regression model. So in recognition memory, no long-term L2 cost was found.

In Chapter 4, some different materials and test types are used to investigate what happens in a cross-lingual condition. In experiments with cued recall of short expository texts and L1-L1, L2-L1 and L2-L2 conditions, no language effects were found. In an experiment with a longer academic text in L2-L1 and L1-L1 a large disadvantage occurs for both multiple choice questions and cued recall, but it seems that the time constraint was too high here. It has been reported that reading or studying in L2 takes longer (for example about 20% in a similar participant population, Dirix et al., submitted). Therefore, we reported this factor and other factors which lead to inconsistent results in this dissertation and other research papers (namely Chen & Donin, 1997; Donin, Graves, & Goyette, 2004; Lee, 1986; Roussel, Joulia, Tricot, & Sweller, 2017). Time constraints, proficiency level, text/test difficulty and language conditions seem to have an impact on the effects (see section 1.2 below).

To find out whether the large L2 recall cost in free recall tasks observed in Chapter 1 is due to production or encoding, the same task is used in a cross-lingual design in Chapter 5, with some necessary adaptations such as vocabulary support and very explicit language instructions. Contrary to our expectation that production in L2 causes the recall cost (at least in free recall), a language effect was found for the L1-to-L1 versus L2-to-L1 condition, but no effect was found for recalling in L1 or L2 after studying in L2. In other words: no effect was found for the language of recall, while an effect of $d = .6$ was observed for the language of encoding.
1.2 Factors involved in L2-costs

Over the course of our empirical research, we observed inconsistencies in our own results and in the literature. Interestingly, task-dependency has been discovered to be crucial to understand effects in other bilingualism research, such as the cognate facilitation effect and homograph inhibition (de Bruijn, Dijkstra, Chwilla, & Schriefers, 2001; Dijkstra, Miwa, Brummelhuis, Sappelli, & Baayen, 2010), but it has barely been touched in the lines of research relevant to the current topic. For the investigation of language-dependent memory in word list and autobiographical memory tasks, a few papers did point out that the way memory is tested affects the way the information is processed and influenced by language. For example, recognition of words benefits from language-congruency, while recall seems to benefit less from this. Francis and Gutiérrez (2012) found that this difference was due to the depth of processing. Older studies point to the same direction: when words are categorised in semantic groups, L1 memory benefits from this categorisation, while L2 memory does not (Nott & Lambert, 1968).

To reach a better understanding of which factors are important and when effects actually arise, we have included relevant studies in Table 2 (a first version of this was presented in Chapter 4). Based on the available test statistics in these studies, we have tried to report the sizes of the effects as well. These effects will be referred to in section 1.4, but first, we will explain which factors seem to be crucial.

First, the test that is used influences the effect. Recognition and recall tests not only tap into a different type of memory, they also result in different levels of difficulty. Within these categories, there are large differences in difficulty as well. The format of a true/false test seems to be so easy that superficial encoding suffices, and that no L2-disadvantage is observed. Nevertheless, this does not mean the content of the statements is necessarily too easy: we have never observed a ceiling effect in the scores. Though recognition memory is tested by multiple choice questions as well, they are much more demanding. More reading is necessary for the test itself, and a good multiple choice test also contains lures that are very probable. As a consequence, understanding of the contents of the text needs to be good. In this sense, the type of test can be considered as a level of difficulty. Apart from recognition tests, recall can also be tested in different ways. Recall is more difficult in itself, but it can contain cues like recognition tests, for example in open and
closed question formats. The most difficult way to retrieve information is when no cue is given whatsoever, as in the free recall task.

Second, Chapter 3 suggests that when time is too limited for a task, this lack of time yields worse results in L2. When the task was easier and ample time was given, we have sometimes observed similar results in L1 and L2 (e.g. Experiment 1 of Chapter 3). This observation accords with findings reported by Chang (2011), who found that the time spent on recall tasks (both the reading and writing phase) in L2 affects the memory scores. In addition, Chen and Donin (1997) found that Chinese-English bilinguals spent 50% more time reading for recall in L2 than in L1 (leading to similar performance). Therefore, research papers should report the time limit and the number of words a text contained to provide full transparency in how demanding a task has been. In addition, when time pressure is too high, students will probably stress out. Stress impairs long-term memory (de Quervain, Roozendaal, Nitsch, McGaugh, & Hock, 2000). When the subjects can use as much time as they want, the time they take should be recorded. This can help to understand whether L2-disadvantages can be compensated by additional effort.

Third, the proficiency level of a subject obviously influences the results of a study (see Chapter 6 and Jeon & Yamashita, 2014). The higher proficiency is in L2, the better the chance of unimpaired higher-level processing. Still, Marian and Fausey (2006) report language-dependency for highly proficient L2 speakers, not for the less proficient group. In addition to this, specific experience in the task might be useful as well. For example, Roussel and colleagues (2017) did not find a disadvantage in computer science students, but they did find it in law student. Perhaps computer students are more trained in English vocabulary since their domain is based largely on English(-only) terminology. Moreover, the task of retrieving information benefits from specific training. If this skill is not automatically transferred to L2 (which is definitely the case for some skills, see Walter, 2007), students who have had English courses before might perform better than students who are new to this situation.

Lastly, the difficulty of the text obviously influences results as well. Not only is reading comprehension more challenging in a difficult text, the risk of encountering unknown (subtechnical) vocabulary in L2 is larger when a text is more difficult. As a consequence, this risk is high for academic texts, which are often opaque and complex.
1.3 Recognition memory

In this dissertation, recognition memory was tested with true/false tests twice. Meanwhile, subsets of those tests were used in two other studies. In a first study (Dirix, Vander Beken, De Bruyne, Brysbaert, & Duyck, submitted), eighty subjects’ eye movements were tracked while they read or studied the texts about the Sun and sea otters (Chapter 2, 3, 5) and excerpts from two academic texts (of which one is used in Chapter 4). Half of the texts were presented in L1, half in L2. Afterwards, they received 10 true/false statements about every text in the language of the text. The aim of the study was to investigate eye movements in L2 studying and the effect of proposition fixation times on the response accuracy for questions about those propositions, but no such effect was found. There was no time limit in this study and subjects on average spent 18% longer on reading or studying in L2 (the study has been cited here a few times for that reason). No difference was found on recognition accuracy in L2 versus L1, which replicates our results found under time constraints. A second study investigated comprehension monitoring in L1 and L2 (Broos, Duyck & Hartsuiker, in preparation). Subjects (N = 68) read the text about the Sun and the sea otters in L1 and L2 (within-subjects), answered 10 true/false statements in the language of the text and estimated how well they understood the texts. In this study, scores were significantly higher in L1 (M = 8.22 out of 10, SD = 1.03) than in L2 (M = 7.89, SD = 0.99; type II Wald chisquare ANOVA χ² = 4.70, p = .03). Since we do not have access to the correlation between the scores in L1 and L2, we cannot calculate the effect size (Cohen’s d must be corrected for the correlation in within-subject designs). The fact that an effect is found here, while no effect was found in four other studies, might have something to do with the selection of questions (only 10, though Dirix et al. also selected only 10). Another possibility is that the effect exists but is very small and only found with large power (N = 68 in a within-subject design with few factors). We cannot check this possibility without the effect size. A last difference might be the reading times of the subjects: perhaps subjects have only shortly read the texts rather than studying them, resulting in the discrepancy with the other results.

As far as recognition memory goes, we can draw two conclusions. In most studies, there is no disadvantage in L2. When a disadvantage is reported, there are indications
that reading time was too short (and the text was perhaps not even read entirely due to time constraints, as in the L2-L1 comparison with L1-L1 in Chapter 3) or we have little information about reading times or effect size. Furthermore, we have found barely any evidence for language-specificity, despite what is reported in studies on the word level and in episodic memory. To our surprise, cross-lingual conditions did not result in lower scores for the true/false test (in Chapter 4). We do consider the possibility that subjects accessed the conceptual level more in this task, since it was preceded by a free recall task, which might have resulted in a testing effect. Altogether, we found surprisingly little evidence for language-dependent recognition memory. Further research should focus on recognition memory for texts (without a preceding recall task) in congruent and incongruent conditions to settle this issue with certainty.

1.4 An overview of effect sizes in bilingual text recall

In Chapter 4, a table was provided containing studies investigating memory for texts and the factors we believe are of importance. To improve our understanding of the size of L2 disadvantages and the conditions in which they appear, a direct comparison between those studies is valuable. Therefore, we have calculated effect sizes of all text recall studies that compared L1 and L2 and that reported enough statistics for us to derive an effect size.

Two popular measures that are often used are Cohen’s $d$ and partial eta squared ($\eta^2_p$), which were also reported in the empirical chapters of this dissertation. As explained by Lakens (2013), they are part of two different ‘families’ of effect sizes. The $d$ family consists of standardized differences between means. The idea behind a standardised effect size is to make results comparable between experiments or studies. The $r$ family ($r^2$, $\eta^2_p$ ...) informs about the strength of association by describing the proportion of variation that is explained by the effect. The choice between those

---

1 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.
measures depends on the viewpoint that is taken. According to some authors, effect size should be generalizable over between- and within-subject designs, while others emphasize the importance of the statistical test, for which it is best to disregard individual differences. To sum up: the families of effect sizes are inherently different because one group takes into account the other factors in the design, while the other group ignores those factors. In addition, measures can be corrected for bias or not. Let us first take a look at the two often reported measures Cohen’s $d$ and $\eta^2_p$.

### 1.4.1 Cohen’s $d$

Cohen (1988) introduced a formula in which the difference between the means of two groups is divided by the standard deviation of the total sample, as in the following formula.

$$ d = \frac{M_1 - M_2}{\sigma} $$

In other words, this measure enables you to compare differences between groups with a standardized value. Lakens (2013) refers to this $d$ as $d_s$ (standardized mean difference for the sample). When the classical Cohen’s $d$ is computed, data is treated as if it came from independent samples. So when the classical measure is used for within-subjects designs, it ignores the experimental set-up. Therefore, when $d$ is calculated for repeated-measures designs (e.g. by Lakens, 2013), it is corrected for the correlation between the paired samples (e.g. $d_{rm}$) or based on a one-sample test of the difference scores ($d_z$). Cohen (1988) proposes a value of $d = .2$ as small, $.5$ as medium and $.8$ as large. To correct for bias in small samples, Lakens advises to use Hedges’ correction when running meta-analyses. He provides the following formula for this correction.

$$ Hedges's \, g_s = Cohen's \, d_s \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right) $$

Durlak (2009) presents the formula $g = \frac{d}{\sqrt{N/df}}$ which results in approximately the same value.
1.4.2 Partial eta squared

In contrast, partial eta squared computes the effect size via sums of squares. The sum of squares of an effect is in the numerator; the denominator contains the sum of squares of the effect and the error sum of squares.

$$\eta^2_p = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

That way, $\eta^2_p$ basically denotes how much variance an effect explains after ruling out other variables. The more factors or covariates are used in an analysis, the more the error sum of squares decreases and $\eta^2_p$ increases, which may lead to an inflation of the measure. An alternative measure which is proposed as the best way to compare between experimental designs is to use generalized eta squared ($\eta^2_G$), which excludes other factors’ variation, resulting in a more comparable measure between studies. Nevertheless, the full ANOVA table is needed to compute $\eta^2_G$ while it is usually not reported in research papers.

1.4.3 A comparison of measures

To sum up the previous sections: Cohen’s $d$ can be a valuable measure to compare effects between studies, though it might simplify effects by ignoring other variables. In addition, the classical Cohen’s $d$ was developed for independent samples. Partial eta squared takes the design of a study into account, but if many covariates are used, $\eta^2_p$ becomes inflated. To get an idea of how these measures are related to each other, we have calculated these two classical measures and the two measures preferred by Lakens (2013) for the L2 recall costs reported in Chapter 2 and Chapter 5. In Chapter 2, a mixed ANOVA was reported including language as a between-subjects factor and test type as a within-subjects factor. Partial eta squared is derived from this ANOVA, and generalised eta squared was calculated from the sums of squares as describe by Lakens (2013). A post-hoc calculation of Cohen’s $d$, was reported based on the results for the free recall test only, in which a language effect was found. In Chapter 5, language was tested within-subjects and a paired samples t-test was reported there (including Cohen’s $d$). To be able to compare all measures, a one-way repeated measures ANOVA was run to derive $\eta^2_p$. 
and $\eta^2_G$ from. In this case, Cohen’s $d_z$ is for a paired sample (and corrected for correlation). In addition, Hedges’ $g$ (corrected for bias) is reported as well, but since sample sizes are large, Hedges’ $g$ is almost equal to Cohen’s $d$.

<table>
<thead>
<tr>
<th>Study</th>
<th>$\eta^2_p$</th>
<th>$\eta^2_G$</th>
<th>Cohen’s $d$</th>
<th>Hedges’ $g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 2 recall cost</td>
<td>.09</td>
<td>.06</td>
<td>.86</td>
<td>.85</td>
</tr>
<tr>
<td>Chapter 5 experiment 1</td>
<td>.27</td>
<td>.09</td>
<td>.61</td>
<td>.60</td>
</tr>
</tbody>
</table>

Table 1. *Comparison of four effect size measures as discussed by Lakens (2013) on two studies reported in this dissertation.*

As the comparison in Table 1 shows, generalised eta squared is smaller than partial eta squared (as it should be based on the formulas), but this difference is very large for the repeated-measure design. It seems that partial eta squared is extremely sensitive to a within-subjects design, overestimating the effect size. If partial eta squared would be calculated for a post-hoc comparison of the relevant groups for Chapter 2 (which is more similar to the information used for Cohen’s $d$), it becomes .16 instead of .09. This shows that partial eta squared takes into account the full design when it is calculated from the F-values in, for example, a mixed ANOVA. When comparing the two studies presented in the table, partial and generalized eta squared are larger for the second study, while Cohen’s $d$ is smaller for that study. To sum up, we argue that partial eta squared should be interpreted with great care, since it seems to be inflated in one-way repeated measures ANOVAs. Generalised eta squared would be more useful, but cannot be calculated from other authors’ research papers. In this sense, Cohen’s $d$ seems to be more useful to compared the effect size in different research designs. In the next section, an effect size table is presented for all relevant studies of L2 text recall providing sufficient information.

1.4.4 Overview of reported language effect sizes

The measures discussed above can be converted to each other (for example in power calculation software such as G*Power; Faul, Erdfelder, Lang, & Buchner, 2007), but they are inherently different and the guidelines for interpreting them were developed based
on the original computations. Therefore, we have chosen to report partial eta squared next to Cohen’s d, keeping as close to the original formulas as possible.

The Cohen’s d reported in our papers with ANOVAs is based on post-hoc calculations of the relevant contrast. 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 contains a condition with L2 text and a translation. For detailed test statistics on all language conditions – so the weaker language plus translation included – with post-hoc comparisons, we refer to the original paper. The partial eta squared is derived from the main language effect in the entire ANOVA, taking additional conditions or factors into account (e.g. test type in our study or all conditions in Roussel et al., 2017). Note that Cohen’s for Chapter 4 was calculated on the cued recall scores only, without the multiple choice-questions that are included in the chapter itself, while the partial eta squared includes the full design and, thus, the multiple-choice questions.

The corrected Hedges’ g would provide an unbiased measure in small sample sizes, but Cohen’s d allows us to calculate confidence intervals for the effect size (and most studies tested large samples). For studies by other authors, Cohen’s d was calculated by using an online effect size calculator (Lenhard & Lenhard, 2016). Cohen’s d effect sizes and their confidence intervals were plotted in Figure 1, to facilitate interpretation.

As can be derived from the table, the differences between studies are consistent: when d is larger for study A than for study B, is also larger in study A than study B. In this table, number of words is no longer reported, but reading time was reported per 100 words. All of the studies that are mentioned used short texts (between 200 and 520 words).
### Table 2. Overview of experiments which directly compared text recall in native and non-native languages, with standardized effect sizes.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Language conditions</th>
<th>L1 and L2</th>
<th>Language factor</th>
<th>Type of recall</th>
<th>Level of proficiency</th>
<th>Min./100 words*</th>
<th>N_{groups}</th>
<th>\eta^2_p</th>
<th>d (CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee, 1986</td>
<td>L2 - L2</td>
<td>L2 - L1</td>
<td>English</td>
<td>Free</td>
<td>1st &amp; 2nd-year FL uni students (no CERF ref)</td>
<td>Self-paced</td>
<td>160</td>
<td>.04**</td>
<td>0.44 (0.22 – 0.66)</td>
</tr>
<tr>
<td>Donin, Graves &amp; Goayette, 2004</td>
<td>L2 - L1</td>
<td>L1 - L1</td>
<td>English</td>
<td>within</td>
<td>Blocked (online)</td>
<td>Intermediate</td>
<td>Self-paced (L2 slower)</td>
<td>16</td>
<td>.74**</td>
</tr>
<tr>
<td>Roussel et al., 2017</td>
<td>L1 - L1 (L2+trans − L1)</td>
<td>L2 - L1</td>
<td>French</td>
<td>between</td>
<td>Cued</td>
<td>B1 (CEFR)</td>
<td>6.22**</td>
<td>34</td>
<td>.27</td>
</tr>
<tr>
<td>Experiment 2*</td>
<td>L1 - L1 (L2+trans − L1)</td>
<td>L2 - L1</td>
<td>French</td>
<td>between</td>
<td>Cued</td>
<td>B1 (CEFR)</td>
<td>6.22</td>
<td>28</td>
<td>.25</td>
</tr>
<tr>
<td>Roussel et al., 2017</td>
<td>L1 - L1</td>
<td>L1 - L1</td>
<td>French</td>
<td>between</td>
<td>Cued</td>
<td>B1 (CEFR)</td>
<td>6.22</td>
<td>36</td>
<td>.02**</td>
</tr>
<tr>
<td>Experiment 3*</td>
<td>L1 - L1 (L2+trans − L1)</td>
<td>L2 - L1</td>
<td>French</td>
<td>between</td>
<td>Cued</td>
<td>B1 (CEFR)</td>
<td>6.22</td>
<td>36</td>
<td>.02**</td>
</tr>
<tr>
<td>Chapter 2</td>
<td>L1 - L1</td>
<td>L2 - L2</td>
<td>Dutch</td>
<td>between</td>
<td>Free</td>
<td>72 (LexTALE) = B2 (CEFR)</td>
<td>2.6</td>
<td>97</td>
<td>.09</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>L1 - L1</td>
<td>L2 - L1</td>
<td>Dutch</td>
<td>within</td>
<td>Cued</td>
<td>72 (LexTALE) = B2 (CEFR)</td>
<td>1.39</td>
<td>53</td>
<td>NA</td>
</tr>
<tr>
<td>Experiment 1A</td>
<td>L1 - L1</td>
<td>L2 - L1</td>
<td>Dutch</td>
<td>within</td>
<td>Cued</td>
<td>72 (LexTALE) = B2 (CEFR)</td>
<td>1.39</td>
<td>63</td>
<td>NA</td>
</tr>
<tr>
<td>Experiment 1B</td>
<td>L2 - L1</td>
<td>L2 - L2</td>
<td>Dutch</td>
<td>between</td>
<td>Cued</td>
<td>72 (LexTALE) = B2 (CEFR)</td>
<td>0.86</td>
<td>62</td>
<td>.07</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>L2 - L1</td>
<td>L1 - L1</td>
<td>Dutch</td>
<td>between</td>
<td>Cued</td>
<td>72 (LexTALE) = B2 (CEFR)</td>
<td>2.6</td>
<td>62</td>
<td>NA</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>L1 - L1</td>
<td>L1 - L1</td>
<td>Dutch</td>
<td>within</td>
<td>Free</td>
<td>72 (LexTALE) = B2 (CEFR)</td>
<td>2.6</td>
<td>59</td>
<td>NA</td>
</tr>
</tbody>
</table>

* 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

* The means are not reported, but more time was spent in L2 reading: F(2,32)= 86.99.

** 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.

** This was not provided in the original paper, but calculated using the F-statistic via the online tool [https://effect-size-calculator.herokuapp.com/#partial-eta-squared-fixed-effects](https://effect-size-calculator.herokuapp.com/#partial-eta-squared-fixed-effects)
More importantly, these calculations illustrate that free recall is not the only task in which clear disadvantages are found. The largest effect size is found in the study by Donin, Graves and Goyette (2004). In that paper, no means and standard deviations are given for the language groups, so effect sizes could only be derived from the F-statistic of the language effect. Subjects in this study had to read short texts that were presented sentence by sentence and recalled the information in blocks (after every fourth sentence). Recall was in the native language, but because the recall was in blocks, this means the L2-L1 condition required the participants to continuously switch between languages. Since bilingualism research has shown switch costs, with stronger inhibition of L1 than of L2, we have some doubts whether this effect size is meaningful.

Other large effects are found in the study by Roussel and colleagues comparing L2-L1 and L1-L1 cued recall. Their third experiment did not reach significance, and though our calculation results in an effect size of Cohen’s $d = .3$, the lower limit of the confidence interval is below zero. Hence, it is unclear whether the effect was just smaller and thus remained undetected with limited power, or whether there is no ‘hidden’ effect. It is
reasonable to believe that participants in these studies performed worse in L2 than ours, since their proficiency level is lower (B2 vs B1 in the CEFR framework). The computer science students in Experiment 3 are probably more trained in using English materials or terminology than the others groups, which made the effect disappear. Next, our free recall study comparing L2-L2 and L1-L1 recall shows a large effect, followed by the L2-L1 and L1-L1 recall study. The cross-lingual study conducted by Lee (1986) is of similar magnitude. Together with the findings reported by Roussel, this clearly shows that encoding language affects memory performance.

Still, several studies (about four) have effects meandering around zero. For Chapter 4, in which cued recall was tested, it can be argued that the effect is smaller than in free recall thanks to the cues. Since we aimed to detect effect sizes of $d = .4$, which is proposed by Ferguson (2009) as the threshold for relevant effects in applied settings such as education, these effects ($d = .1, .2$) were insignificant. Nevertheless, for the choice of foreign language instruction on population level, it can be relevant. Durlak (2009) advocates the importance of effect sizes that are interpreted as small – and sometimes ignored – but of large practical importance. For example, effect sizes of .2 can be of policy interest when they are based on measures for educational achievement (Hedges & Hedberg, 2007, as cited by Durlak, 2009).

The study by Chen and Donin (1997), which has been discussed in previous chapters, did not make the table because too little information was given to derive effect sizes. In that study, L2-L2, L2-L1 and L1-L1 conditions were compared in Chinese-English bilinguals. The authors do not report any significant effects of language, but there seems to be a trend of better recall in L1 compared to L2 after reading a text in L2. Since the study is underpowered, this effect of test language might have remained undetected, but based on the results reported in other studies this trend might be coincidental as well (especially since the subjects might have compensated L2 disadvantages by spending more time on the task). Another interesting remark was made by Chen and Donin. They propose linguistic distance as an explanation for the discrepancy with other studies, in which a production effect had been found (e.g. Lee, 1986). For bilinguals with great linguistic distances between their languages, little overlap in lexical and syntactic representations would possibly tax working memory to a higher extent, interfering with semantic encoding. Though such an ‘overload’ account fits the processing theories we
have mentioned, the fact that lexical and syntactic equivalence would facilitate processing, goes against our view of language-independent semantic memory. The authors add that, to cope with the overload, readers might have avoided deep-level processing, though they were still able to recall comparable amounts of propositions. Though the empirical evidence for superficial encoding is not present, this idea is very close to the levels-of-processing effect that has been mentioned earlier and will be discussed in section 1.5. The linguistic distance can perhaps be added as a factor defining the L2 cost, but most studies reported in our table use languages that are relatively close to each other (English, German, Dutch, French).

In short: though the standardised effect sizes must be handled with caution, our table shows that L2 disadvantages are omnipresent both in cued and uncued text recall. Contrary to what one might intuitively expect, the encoding language definitely plays a crucial role in the quality of recall, though some evidence is also reported for a production cost in L2. In the following paragraphs, we will draw theoretical conclusions about recall.

1.5 Recall and language-dependency

For recall, it is clear that an L2 disadvantage arises. Though experiments with cued recall in this dissertation did not reach significance, a large L2-cost is found in free recall (Chapter 2). We have argued that this is not a production effect but an issue with the information that is encoded (Chapter 5). It might be possible that additional production costs are found if ratings are less lenient towards language mistakes (see results in Roussel et al. 2017; Lee, 1986), but in any case: memory representation of the text is of less quality in L2 than in L1 if the same amount of time is available for encoding.

Does this mean semantic memory at text level is language-dependent? We argue it is not, since the results of an L2-L1 condition are similar to an L2-L2 condition. Language-dependency would yield benefits in congruent conditions versus cross-lingual tasks. These results rather point in the direction of a separate semantic level for information that is shared between languages. In other words: semantic memory is language-independent. Language-dependency which is reported in word list recall and
autobiographical memory is probably limited to episodic memory or induced by specific task requirements.

The disadvantage in L2 in a language-independent memory can be explained by two language-non-selective frameworks that were mentioned in the introduction. The resource hypothesis (Sandoval, Gollan, Ferreira, & Salmon, 2010) assumes that L2 reading requires more working memory capacity for lower-order processes, leaving less capacity for the higher-order process. This is indirectly supported by evidence that lower-order processes are the strongest predictors for reading comprehension in L2, and not in L1 (Jeon & Yamashita, 2014). The weaker-links hypothesis (Finkbeiner, 2002) assumes that L2 representations are less detailed. If we combine the latter idea with the landscape model of Van den Broeck and colleagues (1999), which presents the mental model as a continuously updated result of (cohort co-)activation based on the words/ideas that are read, reading a text in L2 will indeed lead to a weaker mental model. We will further discuss this mental model in section 1.6.

Whether caused by weaker links or by limited resource, L2 processing is more superficial. In this respect, we can understand language differences as a sort of levels-of-processing effect (Craik & Lockhart, 1972). The LOP-effect means that deeper, conceptual processing (primed by a certain task or strategy during studying) results in better memory performance. Francis and Gutiérrez reported that this effect is larger in the native language for word list recognition (2012). The native language benefits more from deeper processing than the non-native language. Since word list recognition is better in L2 than in L1 (because L2-words are less familiar and a list more unique in memory – this is an example of language-dependent memory effects), this basically means that L2 benefitted less of deeper processing than L1. In short, these results show that L2 processing is more superficial, or at least that deeper processing is easier in a native language that in L2. If L2-processing is more superficial, memory tasks requiring deep understanding of a text and active retrieval are far more difficult in L2, which explains the discrepancy between recognition and recall studies reported here. Some evidence for this reasoning can be found in an eye-tracking study with some of the materials used in this dissertation (Dirix et al., submitted). In that study, reading and studying texts in L1 and L2 was compared for short expository texts, without time constraints. In this case, results were similar, but subjects were 20% slower on average in
L2. Interestingly, there were more and longer fixations in L2, while saccades were smaller. This suggests that a lot of attention is directed to the lexical level, resulting in successful word identification and memory, but less successful higher-order processes such as integration of information in the text.

The importance of the depth of processing in L2 reading was investigated by Oded and Walters (2001). They showed that comprehension of an L2 text (measured by comprehension questions) improved after a task that required deeper processing versus a task that required superficial processing. Half of the participants had to summarize the text, the other half had to list details from the text which, according to the authors, distracts from overall comprehension. They found that the first condition indeed improved comprehension and interpreted this as evidence for the levels-of-processing effect in L2. This is in line with our findings. If these participants were tested with a recognition test, their performance might have been better in the listing condition. Apart from influencing the depth of processing, the summary task might also improve results because it is a form of retrieval practice. Retrieval practice has shown to improve memory (Roediger & Butler, 2011).

1.6 The mental model

In the previous sections, we often referred to the mental model of a text, based on the work by Van den Broek and colleagues (1999). Let us take a look at this mental model and the ways in which it is possibly influenced by language.

Barry and Lazarte (1998) describe the mental model as one of three levels of a multilevel representation of texts. The other levels are the verbatim and the textbase representations. The textbase representation is the semantic representation of a text, containing the surface meaning of the text. The mental model is the result of summarizing across propositions or integrating them with prior knowledge, so a representation containing more than only what is in the text. This level is affected by prior knowledge and complexity of the text. Subjects with ‘high knowledge’ make more inferences and elaborations, resulting in a richer and more accurate mental model, at least for complex texts. When texts are easier, it seems as if a rather exact replica of the
textbase is stored. The authors assume a shift takes place in these readers when they are confronted with complex texts, in which they turn to a more top-down process, resulting in a richer mental model and diminished text base. Readers with less knowledge mainly store the propositional information from the text. This high-knowledge-low-knowledge contrast is similar to what we have described happens in L1-L2 contrasts. Readers do not have the language knowledge in L2 which enables them to elaborate or invest in a rich mental model.

This approach looks into the mental model in memory after the reading of a text, the offline product of reading, while others have looked at the online processing of a text (e.g. with eye-tracking techniques). In the tradition of both cognitive science and discourse analysis, Van den Broek and colleagues (1999) developed the Landscape (LS) View of reading to reflect both elements (online and offline) and their bidirectional relations. They argue that reading a text entails understanding what is written and integrating this in prior knowledge. This knowledge is adapted to what is read at the same time. The reading process itself is dynamic in the sense that the mental representation is constantly updated. They list four sources of activation: the text fragment being read in the current reading ‘cycle’, the information from the previous cycle, reactivation of previous concepts, and prior knowledge. Since limited resources are available for processing all of the information, this results in a continuously fluctuating activation with ‘peaks’ and ‘valleys’ for concepts during different cycles (hence the term landscape).

Keeping this view in mind, it is probable that the model of an L2 text is of less quality. If less working memory capacity is left (according to the resource hypothesis, see previous section), than there is less room for information from these four sources or for the interaction between them. If L2-representations are less detailed (as presumed in the weaker-links hypothesis, see previous section), less previously encountered concepts might be reactivated (because semantic association between them is weaker) or less related prior knowledge comes to mind. In addition, when the meaning of concepts in the texts is not known, no interactions happen at all. The fact that high knowledge L2 readers (Larry and Bazarte, 1998) create a richer model supports these ideas. High knowledge readers will have more prior knowledge and more semantic cohort activation, which makes up for the L2 disadvantage to some extent. They seem to rely more on this knowledge or try to compose a coherent mental representation of a text more actively.
when the syntactic complexity increases, while easier texts are just stored in a more text-based manner. In other words: when sufficient resources are available (prior knowledge, known vocabulary, etc) and when the structure of the text is less straight-forward, readers engage in the creation of an abstract mental model of the text. When these resources are not available, readers stay closer to verbatim or textbase representation (resulting in equal recognition scores but hampered recall).

To conclude: the current dissertation provides corroborating evidence of L2-disadvantages in semantic memory, which is probably language-independent, but hampered in deeper processing tasks. Language non-selective frameworks such as the resource hypothesis and the weaker-links hypothesis can account for this. The nuance that recognition memory is not or less hampered than recall is explained by the levels-of-processing effect.

2. Implications for education

It is clear that the use of a non-native language such as English for education entails a few risks. Still, the fact that no additional forgetting is found (in recognition memory) is good news for the knowledge of students who receive foreign language instruction. EMI need not be an issue per se, as long as the following risks are taking into account².

1) **Receptive knowledge can be underestimated when L2 recall tests are used:**

If the goal of a course or test is for students to detect false statements, recognise details, or to evaluate conclusions that were instructed, they probably perform as well in L2 as in L1. But they are not able to express this knowledge independently (without a

---

² Some of these directions have been discussed in more detail by De Bruyne and Vander Beken (submitted) in a short Dutch handbook text.
cue). In other words: when they have to recall this information, they will not be able to display the receptive or marginal knowledge they have.

2) **Knowledge can be overestimated in L2 recognition tests:**

The opposite of the previous argument is true as well. If students have to take simple recognition tests, they perform at an equal level in L2, but they will not be able to use that same information in an active way. That is, writing a recall protocol will be more difficult and their understanding seems to be less ‘deep’.

3) **Studying in a non-native language is more time-consuming:**

In the current dissertation, time limits were used for all tests. Since we know from other studies (e.g. Dirix et al., submitted) that subjects take more time to read or study the same materials in L2 (e.g. 18% in Dirix et al.), it is possible that the costs that were reported here can be compensated if subjects put more time and effort in the task. In that case, English-taught programmes or English handbooks will be more challenging and time-consuming for students. It seems that teaching an English course also takes more time (Van Mol & Valcke, 2017), both in preparation and during the course itself.

4) **Presenting tests in a native language does not solve the L2 disadvantage:**

Especially in Chapter 5, it has become clear that the recall cost is not merely a production cost, but mostly an issue with encoding or storage of an L2 text. For very demanding tasks, recall suffers from studying in L2. As a consequence, the risks of EMI are not entirely solved by providing the opportunity to take exams in a native language.

5) **Vocabulary support is crucial:**

From a theoretical point of view, it seems that the mental model is impaired mostly by unknown vocabulary and vocabulary that is misunderstood (e.g. false friends) or only
partially understood. In addition, other researchers have argued that academic or subtechnical vocabulary is crucial for the understanding of academic texts (Nation, 2013) or of definitions of domain-specific vocabulary (Gablasova, 2015), while this is often not instructed explicitly. This issue should receive attention when a foreign language is introduced for instruction.

6) No additional care seems to be necessary for long-term memory:

If the results from recognition memory hold in recall tests, then it seems that long-term storage of information in L2 is not problematic in itself. Of course, weak mental models will decay sooner, but for now, we can assume that a model which suffices for a certain test in L2 will suffice for that same test after a few weeks. Hence, when the initial understanding is improved, disadvantages are probably avoided on the long term.

3. Limitations and future directions

Many questions remain to be answered, although the current dissertation has provided a first schematic overview of certain effects. Of course, there are some limitations to the research reported in this dissertation, but these often go hand in hand with ecological validity. In the next paragraphs, some limitations and suggestions for future research are discussed.

3.1 Limitations

The participant pool used here was very similar along all experiments, both in language level and in background knowledge. All participants were Dutch-English
bilinguals from Flanders with a high-intermediate level of English proficiency corresponding to the B2-level in the Common European Framework. As such, we cannot generalise these findings to other bilingual populations with a different L1 and a different proficiency level in L2, or to balanced bilinguals. Still, this participant pool is a good proxy for the population of university students studying materials in a second language, which is one of the main concerns of this topic. As a consequence, practical implications are important. Furthermore, since many variables are involved in the results, it is difficult to predict performance in similar situations. In this respect, the overview table is very informative. A third threat for generalisability is the fact that limited materials are used. It is not feasible to test dozens of stimuli, like in experiments on the word-level, when texts and tests are the stimuli (note that in Chapter 2, questions were included as random effects and thus accounted for). As a consequence, the effects reported in this dissertation might be defined by certain features of the texts (e.g. the amount of domain-specific vocabulary). In addition, it has not been tested what happens when subjects need to study a longer text, such as an entire chapter or a full academic article. Most studies use short texts, such as the 300-words texts in this dissertation. Therefore, cumulative evidence from studies with other and longer texts is important to check whether our results hold.

In a broader sense, the research was limited to studying from text. Internationalisation of education also leads to the use of foreign languages for the teaching of entire courses. On a theoretical level as well, the studies reported here cannot increase our understanding of the auditory modality. Some research has been done into L2 listening comprehension. For example, Marian and Fausey (2006) found evidence for language-specificity in the auditory modality: bilinguals recalled a lecture better in the congruent language than in the other language. Reithofer (2013) investigated whether a conference talk presented by a foreign language speaker or a simultaneous interpretation of the same talk resulted in the best cognitive end-result in the audience. She found that interpretation led to better memory and she therefore concludes that the use of English as a (foreign) lingua franca is detrimental. Nevertheless, we think other elements in that study might explain the difference. For example, the simultaneous interpretation was edited before it was played to the audience, and might have been less complex in the first place. To sum up: our results cannot predict what happens to bilingual memory in the
auditory modality, and the studies that are available, are not very informative about foreign language instruction yet.

3.2 Future directions

The data used in this dissertation leave room for further analytical work. For example, qualitative analyses of the free recall protocols could provide information on the mental model in L2 by looking into the types of errors and the information which is or is not remembered. Many studies investigated recall protocols qualitatively or used think-aloud protocols to get a grasp of the mental processes behind reading comprehension, but a direct qualitative comparison of the output in L1 and L2 could really improve our understanding of the L2 disadvantages. In addition, ideas were not weighted in the current dissertation, while the recall of some ideas might be more crucial to performance in real-life situations. Perhaps the type of ideas remembered in L1 and L2 differs.

Furthermore, it would be interesting to investigate the effect of production in detail. When do these effects arise? Are the observed effects the consequence of punishing language mistakes? How much of the performance is still graded on a memory test if language mistakes are punished?

In relation to chapter 3, it would also be valuable to investigate forgetting in L2 for other tasks than true/false tests. Up to this point, we have no clue what happens to recall in the long term: does more knowledge become marginal in L2? Is the proportion of forgetting related to the L2-disadvantage?

As we have mentioned in the previous section, a limited number of texts and only one type of bilinguals was tested. The state of the art would improve spectacularly if labs could work together on an international level, running each other’s experiments in their participant pool. This would immediately show how the effects (or effect sizes) relate to each other.

Lastly, apart from all these directions to clarify what has been found so far, another line of research could focus on the compensation for L2 disadvantages. While we deliberately kept the conditions exactly the same in L1 and L2 to avoid for compensation,
there are reasons to believe that performance improves in L2 if more time and effort is spent on the task. Does the effect disappear if more time is provided (20% more, for example, see Dirix et al., submitted)? Do eye movements change in later stages of reading; i.e. do they become more L1-like once the initial word processing was successful? How can deeper processing be encouraged in L2 (Oded and Walter), and what support do students need in particular?

3.3 A need for theoretical approaches

It is striking to see that not a single theoretical model seems to exist for this line of research, in contrast to the large number of models on the word level. Some theories are available for L2 reading comprehension and for mental models in general, but few on memory for higher-order processes in L2. Though it is obviously more complex to develop a model for higher-order processes than for lower-order processes, we need some sort of model to predict L2 performance. Such a model could be inspired by reading comprehension models like DIME (Cromley and Azevedo, 2007), but with adapted weight for factors that are more important for memory, and with specific predictors for L2.

Despite the lack of such a model, some careful conclusions can be made from this dissertation. On the one hand, no L2 disadvantage arises on most recognition tests, even under time limit. On the other hand, there is quite some evidence that L2 recall is costly, at least for subjects with lower proficiency levels and in a rather difficult task. This cost arises during the encoding or storage process, a process which seems to be more superficial in L2 than in L1, due to a taxed working memory or weaker links. We argue that semantic memory for texts (or the mental model) is language-independent, since no evidence for language-specificity was found.
References


Theory and background

This dissertation reports on an investigation of memory for texts in a non-native language. In a series of controlled experimental set-ups, two main research questions are answered. The first and rather practical question is whether any disadvantages occur when English as a non-native language is used as a medium of instruction. The second question is about how memory for texts is structured in bilinguals. Is memory for texts language-dependent or is the information stored in an abstract manner? In other words, is semantic memory language-dependent? Little research has been done into the latter question, at least at the text level.

English as a medium of Instruction

About half of the world’s population is bilingual, in the sense that they know more than one language to some extent (Grosjean, 1989). With increasing globalisation and internationalisation of higher education, English is growing as a world language and increasingly used as a medium of instruction in Europe (Wächter & Maiworm, 2014) and worldwide (Doiz, Lasagabaster, & Sierra, 2013). In the Netherlands and Flanders (Belgium), English is used to a different extent. The Netherlands are ranked as the European leader in providing English-taught programmes, while Belgium is on rank 17 in the same report (Wächter & Maiworm, 2014). In concrete numbers, twenty percent of bachelor degrees in the Netherlands are offered in English, with another 10 percent offering the choice between Dutch and English. For master degrees, this is 59% (note that these numbers vary greatly depending on the subject), and for colleges (“hogescholen”) it is 8% (KNAW, 2017). In Flanders, this was only 22% for masters and under 2% for bachelor degrees in 2015 (VLOR, 2017), and it is limited to a maximum of 35% and 6% respectively by law. In addition to following courses taught in English, some Dutch studies also require the use of English handbooks (e.g. Dirix et al., 2017). Despite the different size of the phenomenon, the use of English is heavily debated in both countries. This debate pivots mostly around ideological issues and contains some legitimate arguments of course, but little is known about the cognitive consequences for students studying new materials in a foreign language.
Bilingual memory

A large body of bilingualism research has investigated semantic memory. Semantic memory is part of our declarative or explicit knowledge – the knowledge about the world we can verbalize – and contains our factual or world knowledge which is not tied to a specific event (Hardt, Nader, & Nadel, 2013), as opposed to episodic memory which contains these specific events. Since this research has mainly focused on the word level, semantic memory in this case refers to the meaning of words in the mental lexicon. Though different models for word recognition are available, researchers have reached a consensus that the meaning of words is represented at an abstract level, separate from lexical representations and shared between languages (the connections between those levels depend on the model). In addition, they agree that lexical access is language non-selective, since cross-lingual interference has been observed even in sentence reading (e.g. Duyck, Assche, Drieghe, & Hartsuiker, 2007). In the same vein, three frameworks from the language non-selective access view predict that L2 reading is hampered, but they assume different processes behind hampered L2 processing.

A first account, the CROSS-LINGUISTIC INTERFERENCE HYPOTHESIS (Weber & Cutler, 2004), assumes competition between representations of both languages at the lexical level which interferes with recognition. When reading a text in a second language, this interference would hinder and slow down the encoding process, which is what one would intuitively expect. Another possible reason for slower and hampered L2 reading is given in the WEAKER-LINKS HYPOTHESIS. This framework predicts the same issues for L2 words as for low-frequent L1 words. Those words are less familiar and their semantic representations are less detailed (Finkbeiner, 2002). The third and last hypothesis assigns L2 disadvantages to a different prerequisite of reading: working memory capacity. The RESOURCE HYPOTHESIS presumes that working memory is taken up more in a second language, leaving less capacity for higher order processing such as integrating prior and new information and monitoring comprehension (Sandoval, Gollan, Ferreira, & Salmon, 2010).

Based on these frameworks, one would hypothesize that information learnt from a text is also stored at an abstract language-independent level, but that the language of input (L2) hampers encoding. As put by Oakhill, Cain and Elbro, “what readers remember of a text is not the wording [...] but the meaning” (2014, p. 11). The idea that a
text is not stored in a verbatim manner corresponds to an idea in older studies on text and discourse representation (Alba & Hasher, 1983; Schank, 1972). Because readers remember the gist of the text instead of the wordings, Schank (1980) concluded that meaning is represented free of language.

Still, there is an entirely different strand of research which points to language-dependency in memory. The ENCODING-SPECIFICITY PRINCIPLE (Tulving & Thomson, 1973) states that more information is recalled when the context of encoding and retrieval are similar. For example, when learning a text in noisy or silent conditions, you remember more of the text in the same (context-congruent) noise-condition (Grant et al., 1998). Language also serves as such a context, as has been shown in three types of memory. Firstly, autobiographical memory research has shown that people recall earlier, more and more detailed memories in the language the events took place in (Marian & Neisser, 2000). Secondly, word list recognition and recall seems to benefit from language-congruency (Watkins & Peynircioglu, 1983). Thirdly, listening comprehension also results in better recall performance in the language of listening (Marian & Fausey, 2006). This evidence contrasts the idea that meaning is stored entirely independent of language, and even shows advantages for L2.

The current dissertation

When investigating memory for text, a few things need to be taken into account. Firstly reading comprehension is automatically involved: a reader can only remember from a text what they understand in the first place. Hence, we are also indirectly measuring reading comprehension. To avoid language effects in between-subject design actually being group differences in reading comprehension, we measured some variables that are predictive of reading comprehension, such as working memory (Seigneuric, Ehrlich, Oakhill, & Yuill, 2000) and proficiency (Jeon & Yamashita, 2014). In addition, we chose to administer test after participants have studied a text, without text accessibility. That way, we measure memory of a text rather than its comprehension. Secondly, memory can be measured or operationalised in different ways. Recognition and recall are very different processes: active retrieval is only necessary for the latter.
Recognition tests thus tap into something that is called **marginal knowledge** (Cantor, Eslick, Marsh, Bjork, & Bjork, 2014): information we possess but cannot actively retrieve. Both types of remembering can be split up further in cued versus uncued recall and different recognition tests such as true/false judgment and multiple choice tests.

**In Chapter 2**, both recognition and recall of texts were tested in a native versus a non-native language. Two language groups received a short expository text and a subsequent test in either Dutch or English. The test could be a true/false judgement test (recognition) or a free recall task, in which it is asked to write down everything you remember from a text. Results show no language difference on the true/false test, but a large L2 disadvantage on the free recall task (Cohen’s d = .86). Two possible explanations for this finding are given. Either production in L2 is hampered to such an extent that subjects had difficulty expressing their knowledge (although language mistakes were not punished), or the mental model of the text is of lesser quality, resulting in successful recognition but worse recall. The latter could be the result of more superficial processing or of weaker links in L2, resulting in less (cohort) activation of the represented concepts and weaker memory traces, and is in line with the **landscape model** of reading (van den Broek, Young, Tzeng, & Linderholm, 1999). If this were true, recognition memory, which is sufficient for an immediate test, should decay sooner in L2 than in L1, because weaker memory traces also decay sooner.

Therefore, we tested recognition memory on an immediate and delayed test in **Chapter 3**, using the same materials. The delayed test was administered either after a day, a week, or a month. The forgetting curves look very similar in L1 and L2 (with absolute scores slightly lower in L2) and no language effect was found in a mixed logistic regression model. This points to the direction that production is the key element to the large L2 recall cost.

In **Chapter 4**, some different materials and test types are used to investigate what happens in a cross-lingual condition. If production is the issue, the disadvantage should disappear in an L2-to-L1 condition. In experiments with cued recall of short expository texts and L1-L1, L2-L1 and L2-L2 conditions, no language effects were found. In an experiment with a longer academic text in L2-L1 and L1-L1 a large disadvantage occurs for both multiple choice questions and cued recall, but it seems that the time constraint was too high here. It has been reported that reading or studying in L2 takes longer (for
example about 20% in a similar participant population, Dirix et al., submitted). Therefore, we reported this factor and other factors which lead to inconsistent results in this dissertation and other research papers (Chen & Donin, 1997; Donin, Graves, & Goyette, 2004; Lee, 1986; Roussel, Joulia, Tricot, & Sweller, 2017). Time constraints, proficiency level, text/test difficulty and language conditions seem to have an impact on the effects.

To find out whether the large L2 recall cost in free recall tasks is due to production or encoding, the same task is used in a cross-lingual design in Chapter 5, with some necessary adaptations such as vocabulary support and very explicit language instructions. Contrary to our expectation that production in L2 causes the recall cost (at least in free recall), a language effect was found for the L1-to-L1 versus L2-to-L1 condition, but no effect was found for recalling in L1 or L2 after studying in L2. In other words: the encoding language resulted in a large effect (d = .6), the production language had no effect. This led to the conclusion that hampered or superficial encoding in L2 leads to a weaker mental model of the text, which results in difficulties in uncued retrieval.

In addition, Chapter 6 reports some exploratory analyses on motivation and attitudes of students towards English as a medium of Instruction, showing that motivation for reading is lower in L2 than in L1, but that in general, students have positive opinions about EMI.

Chapter 7, the general discussion, goes into the theoretical consequences of these findings. Since the cross-lingual conditions in Chapter 4 and Chapter 5 do not result in worse performance, no evidence is found for language-dependency in semantic memory for texts. The results can be explained by the resource hypothesis or the weaker-links hypothesis and by superficial processing in L2 in line with the LEVELS-OF-PROCESSING EFFECT (Craik & Lockhart, 1972; see also Francis & Gutiérrez, 2012). This means that bilingual memory for texts fits in the language non-selective access view with an abstract language-independent level for meaning.
References


Theorie en achtergrond

Dit proefschrift rapporteert onderzoek naar het geheugen voor teksten in een niet-moedertaal. In een reeks gecontroleerde experimenten werden twee belangrijke onderzoeksvragen beantwoord. De eerste, eerder praktische vraag is of er nadelen of problemen opduiken wanneer Engels wordt gebruikt als instructietaal bij mensen met een andere moedertaal. De tweede, meer theoretische vraag gaat over welke vorm het geheugen voor teksten heeft bij tweetaligen. Is het geheugen voor teksten taalafhankelijk of wordt de informatie op een abstracte manier opgeslagen? Met andere woorden: is het semantisch geheugen taalafhankelijk? Er bestaat tot nog toe weinig onderzoek naar de laatste vraag, tenminste op tekstniveau.

Engels als instructietaal

Ongeveer de helft van de wereldbevolking is tweetalig in de zin dat ze van meer dan één taal enige kennis hebben (Grosjean, 1989). Met de groeiende globalisering en de internationalisering van het hoger onderwijs groeit Engels als wereldtaal en wordt het meer en meer gebruikt als instructietaal in Europa (Wächter & Maiworm, 2014) en wereldwijd (Doiz, Lasagabaster, & Sierra, 2013). In Nederland en Vlaanderen wordt Engels in verschillende mate gebruikt als instructietaal. Nederland wordt als Europees leider beschouwd wat betreft het voorzien in Engelstalige studieprogramma’s, terwijl België in hetzelfde rapport op plaats 17 staat (Wächter & Maiworm, 2014). In concrete getallen: twintig percent van de bacheloropleidingen in Nederland wordt in het Engels aangeboden, met nog een extra 10 procent waarbij men de keuze krijgt tussen Nederlands en Engels. Voor masteropleidingen bedraagt dit 59% (merk op dat deze cijfers sterk afhankelijk zijn van de studierichting), en voor hogescholen 8% (KNAW, 2017). In Vlaanderen was dit slechts 22% voor de masters en minder dan 2% voor bacheloropleidingen in 2015 (VLOR, 2017) en is dit bij wet gelimiteerd tot maximum 35% en 6%. Bovendien wordt, behalve colleges in het Engels volgen, ook voor sommige Nederlandstalige vakken de studie van Engelse handboeken verwacht (e.g. Dirix et al., 2017). Hoewel het fenomeen van een verschillende grootteorde is, is het gebruik van Engels in het hoger onderwijs in beide landen onderwerp van hevige discussie. Deze
discussie draait voornamelijk rond ideologische kwesties en bevat uiteraard legitieme
argumenten, maar er is weinig geweten over de cognitieve gevolgen voor studenten die
nieuwe leerstof in een vreemde taal studeren.

*Het tweetalig geheugen*

Een groot aantal studies in de literatuurtraditie rond tweetaligheid hebben het
semantisch geheugen onderzocht. Het semantisch geheugen wordt onderverdeeld onder
declaratieve of expliciete kennis – de kennis over de wereld die we onder woorden
c kunnen brengen – en bevat onze feitelijke of wereldkennis die niet verbonden is aan een
specifieke gebeurtenis (Hardt, Nader, & Nadel, 2013), in tegenstelling tot het episodisch
geheugen dat specifieke gebeurtenissen bevat. Omdat die studies zich voornamelijk
concentreerden op het woordniveau, duidt het semantisch geheugen in dit geval op de
betekenis van woorden in het mentale lexicon. Hoewel er uiteenlopende theoretische
modellen voor woordherkenning bestaan, zijn onderzoekers het erover eens dat de
betekenis van woorden op een abstract niveau geregerepresenteerd is, afzonderlijk van
lexicale representaties en gedeeld tussen talen (de connecties tussen de niveaus hangen
dan weer af van het model). Daarbovenop is er een consensus dat lexicale toegang niet
taalselectief is, omdat interlinguale interferentie is geobserveerd, zelfs tijdens het lezen
van zinnen (Duyck, Assche, Drieghe, & Hartsuiker, 2007). Binnen dezelfde achtergrond
voorspellen drie theoretische kaders uit de niet- taal-selectieve visie dat het leesproces in
L2 belemmerd is, maar ze gaan uit van verschillende processen als oorzaak van dat
gehinderde proces.

Een eerste verklaring, de *CROSS-LINGUISTIC INTERFERE-NCE HYPOTHESE* (Weber &
Cutler, 2004), gaat uit van competitie tussen representaties van beide talen op het
lexicale niveau die interfererert met herkenning. Wanneer iemand een tekst in een tweede
taal leest, zou deze interferentie het encoderingsproces verstoren en vertragen, iets wat
we ook intuitief zouden verwachten. Een andere mogelijke verklaring voor trager en
belemmerd lezen in L2 wordt gegeven door de *WEAKER-LINKS HYPOTHESE*. Deze theorie
voorspelt dezelfde problemen voor L2-woorden als voor laagfrequente woorden in L1.
Die woorden zijn minder vertrouwd en hun semantische representaties minder
gedetailleerd (Finkbeiner, 2002). De derde en laatste hypothese schrijft L2-nadelen toe
aan een andere noodzakelijke voorwaarde voor lezen: de capaciteit van het werkgeheugen. Deze *RESOURCE HYPOTHESIS* veronderstelt dat het werkgeheugen meer in beslag wordt genomen in een tweede taal, waardoor er minder capaciteit overblijft voor processen van een hogere orde zoals de integratie van voorkennis en nieuwe kennis en het monitoren van tekstbegrip (Sandoval, Gollan, Ferreira, & Salmon, 2010).


Het onderzoek in dit proefschrift


In Hoofdstuk 2 werden zowel herkenning als herinnering van teksten getest in een moedertaal en een niet-moedertaal. Twee taalgroepen kregen een korte informatieve tekst en vervolgens een test ofwel in het Nederlands, ofwel in het Engels. De test kon een waar/niet waar-test (herkenning) of een vrije herinneringstaak zijn (uncued), waarin wordt gevraagd alles op te schrijven wat je je herinnert uit een tekst. De resultaten tonen geen taalverschil op de waar/niet waar-test, maar een groot nadeel in L2 op de vrije herinneringstaak (met een effect size van Cohens d = .86). Er worden twee mogelijke verklaringen gegeven. Ofwel is de productie in L2 zodanig belemmerd dat de proefpersonen moeilijkheden hadden om hun kennis uit te drukken (hoewel taalfouten niet bestraft werden), ofwel is het mentale model van een tekst van slechtere kwaliteit, wat resulteert in succesvolle herkenning maar slechtere herinnering. Dat zou het gevolg kunnen zijn van oppervlakkigere verwerking of zwakkere verbindingen met de representatie (weaker links) in L2, resulterend in verminderde (cohort) activatie van de
gerepresenteerde concepten en zwakkere geheugensporen en komt overeen met het landschapsmodel voor lezen (van den Broek, Young, Tzeng, & Linderholm, 1999). Als dit klopt zou herkenning in L2 sneller moeten achteruit gaan dan in L1, want zwakkere geheugensporen vervagen ook sneller.

Daarom testten we herkenning met een onmiddellijke en een uitgestelde test (beschreven in Hoofdstuk 3) met dezelfde materialen. De uitgestelde test werd afgenomen na een dag, een week of een maand. De vergeetcurves zagen er erg gelijkaardig uit in L1 en L2 (met absolute scores iets lager in L2) en er werd geen taaleffect gevonden in een mixed logistisch regressiemodel. Dit duidt erop dat productie het cruciale element is dat bijdraagt tot de grote L2 herinneringskost.

In Hoofdstuk 4 worden andere materialen en testtypes gebruikt om te onderzoeken wat er gebeurt in een taaloverschrijdende conditie. Als productie het probleem is, zou het nadeel moeten verdwijnen in een L2-naar-L1-conditie. In experimenten met cued recall (open of gesloten vragen) van korte informatieve teksten in L1-L1-, L2-L1- en L2-L1-condities werd er geen significant taaleffect gevonden. In een experiment met een langere academische tekst in L2-L1- en L1-L1-condities was er wel een groot nadeel voor zowel meerkeuzevragen als open/gesloten vragen, maar dit is waarschijnlijk te wijten aan een te strenge tijdslimiet. Het is namelijk gerapporteerd dat lezen of studeren in L2 meer tijd vraagt (bijvoorbeeld 20% in een gelijkaardige groep studenten, Dirix et al., ingestuurd). Daarom lijstten we deze en andere factoren op die tot inconsistente resultaten hebben geleid in dit proefschrift en andere studies (Chen & Donin, 1997; Donin, Graves, & Goyette, 2004; Lee, 1986; Roussel, Joulia, Tricot, & Sweller, 2017). Tijdsrestricties, taalvaardigheidsniveau, moeilijkheidsgraad van de tekst/test en taalcondities lijken een impact te hebben op de effecten.

Om te achterhalen of de grote herinneringskost in L2 te wijten is aan productie of encodering, wordt dezelfde taak gebruikt in een taaloverschrijdend design in Hoofdstuk 5, met enkele noodzakelijke aanpassingen zoals woordenschatondersteuning en expliciete instructies over de taalcondities. In tegenstelling tot onze verwachting dat L2-productie de herinneringskost veroorzaakt (tenminste in vrije herinnering), werd er een taaleffect gevonden voor de L1-naar-L1 versus de L2-naar-L1-conditie, terwijl er geen effect werd gevonden voor het herinneren in L1 of L2 na studeren in L2. Met andere woorden: de taal van het studeren resulteerde in een groot effect (d = .6); de taal van de
taak vertoonde geen effect. Dit leidde tot de conclusie dat het belemmerde of oppervlakkige encoderen in L2 leidt tot een zwakker mentaal model van de tekst, wat resulteert in moeilijkheden bij herinneren zonder aanwijzingen.

Daarnaast rapporteert **Hoofdstuk 6** nog de motivatie en attitudes van studenten ten opzichte van Engels als instructiestaal en toont zo dat de motivatie voor lezen lager is in L2 dan in L1, maar dat studenten over het algemeen wel een positieve mening hebben over Engels als instructiestaal.

**Hoofdstuk 7**, de algemene discussie, bespreekt de theoretische consequenties van deze bevindingen. Gezien de taaloverschrijdende condities in Hoofdstuk 4 en Hoofdstuk 5 geen slechtere prestaties opleveren, is er geen evidentie gevonden voor taalafhankelijkheid in het semantisch geheugen voor teksten. De resultaten kunnen worden verklaard door de resource hypothesis of de weaker-links hypothese en door oppervlakkige verwerking in L2, wat ook overeenkomt met het effect van verwerkingsniveaus (levels-of-processing effect, Craik & Lockhart, 1972; see also Francis & Gutiérrez, 2012). Dit betekent dat het tweetalig geheugen voor teksten past in de niet-taalselectieve toegangsvisie met een abstract taalonafhankelijk niveau voor betekenis.
Referenties


APPENDICES:

TEXTS AND TEST MATERIALS
Appendix A: biology text/tests (Chapter 2, 3 and 5)

Appendix A.1 Biology texts used in Chapter 2, 3 and 5

Appendix A.1.1: Texts about The Sun and the Sea Otters

Based on:


Materials received from the authors for research purposes.
The Sun

The Sun today is a yellow dwarf star. It is fueled by thermonuclear reactions near its center that convert hydrogen to helium. The Sun has existed in its present state for about 4 billion, 600 million years and is thousands of times larger than the Earth.

By studying other stars, astronomers can predict what the rest of the Sun’s life will be like. About 5 billion years from now, the core of the Sun will shrink and become hotter. The surface temperature will fall. The higher temperature of the center will increase the rate of thermonuclear reactions. The outer regions of the Sun will expand approximately 55 million kilometers, which is about the distance to Mercury. The Sun will then be a red giant star. Temperatures on the Earth will become too hot for life to exist.

Once the Sun has used up its thermonuclear energy as a red giant, it will begin to shrink. After it shrinks to the size of the Earth, it will become a white dwarf star. The Sun may throw off huge amounts of gases in violent eruptions called nova explosions as it changes from a red giant to a white dwarf.

After billions of years as a white dwarf, the Sun will have used up all its fuel and will have lost its heat. Such a star is called a black dwarf. After the sun has become a black dwarf, the Earth will be dark and cold. If any atmosphere remains there it will have frozen onto the Earth’s surface.

De zon

De zon is vandaag een gele dwergster. Ze wordt gevoed door thermonucleaire reacties rond het centrum die waterstof omzetten naar helium. De zon bestaat in haar huidige toestand ongeveer 4 miljard en 600 miljoen jaar en is duizenden keren groter dan de aarde.


Eens de zon als een rode reus alle thermonucleaire energie opgebruikt heeft, zal ze beginnen krimpen. Nadat ze gekrompen is tot de grootte van de aarde, zal ze een witte dwergster worden. Wanneer de zon verandert van een rode reus naar een witte dwerg kan ze, in gewelddadige uitbarstingen die we nova-explosies noemen, grote hoeveelheden gas afscheiden.

Na miljarden jaren als een witte dwerg zal de zon al haar brandstof opgebruikt hebben en zal ze haar warmte verliezen. Zo’n ster noemen we een zwarte dwerg. Nadat de zon een zwarte dwerg geworden is, zal de aarde donker en koud zijn. Als er daar een atmosfeer overblijft, zal die vastgevroren zijn aan het aardoppervlak.
Sea Otters

Sea otters dwell in the North Pacific. They are the largest of the mustelids, a group that also includes freshwater otters, weasels, and badgers. They are from 1 to 1.5 meters long and most weigh from 25 to 40 kilograms. Large males may weigh 45 kilograms or more.

Unlike most marine mammals, such as seals or dolphins, sea otters lack a layer of blubber, and therefore have to eat up to 30 percent of their body weight a day in clams, crabs, fish, octopus, squids, and other delicacies to maintain body heat. Their voracious appetites do not create food shortages, however, because they are picky eaters, each animal preferring only a few food types. Thus no single type of food source is exhausted. Sea otters play an important environmental role by protecting forests of seaweed called kelp, which provide shelter and nutrients to many species. Certain sea otters feast on invertebrates, like sea urchins and abalones, that destroy kelp.

Sea otters eat and sleep while floating on their backs, often on masses of kelp. They seldom come on shore. Sea otters keep warm by means of their luxuriant double-layered fur, the densest among animals. The soft outer fur forms a protective cover that keeps the fine underfur dry. A couple of square centimeters of underfur contain up to one million hairs. Unfortunately, this essential feature almost led to their extinction, as commercial hunters drastically reduced their numbers.

Under government protection, the sea otter population has recovered. However, occasionally unfortunate events have damaged the sea otter population. For example, in 1989, up to 5000 sea otters perished when the Exxon Valdez spilled oil in Prince William Sound, Alaska.
Zeeotters

Zeeotters vertoeven in de Stille Oceaan. Zij zijn de grootste onder de marterachtigen, een groep die ook uit zoetwaterotters, wezels, en dassen bestaat. Ze zijn 1 tot 1,5 meter lang, en de meeste wegen tussen de 25 en 40 kilogram. Grote mannetjes kunnen 45 kilogram of meer wegen.

In tegenstelling tot de meeste zeezoogdieren, zoals zeehonden of dolfijnen, hebben zeeotters geen speklaag, waardoor ze per dag tot 30 procent van hun lichaamsgewicht aan mosselen, krab, vis, octopus, inktvis, en andere delicatessen moeten eten om hun lichaamswarmte te behouden. Hun gulzige eetlust veroorzaakt echter geen voedseltekorten, omdat ze kieskeurige eters zijn; elk dier verkliest slechts een aantal voedseltypes. Zo wordt geen enkele soort van voedselbron uitgeput. Zeeotters spelen een belangrijke rol in het milieu door de bescherming van wouden van een zeewier dat kelp heet, en dat een schuilplaats en voedingsstoffen voorziet voor vele diersoorten. Bepaalde zeeotters verslinden ongewervelde dieren, zoals zee-egels en zeeoren, die kelp vernietigen.

Zeeotters eten en slapen terwijl ze op hun rug drijven, vaak op kelpmassa’s. Ze komen zelden op de kust. Zeeotters houden zich warm door middel van hun luxueuze dubbelgelaagde vacht, de dikste onder de dieren. De zachte buitenvacht vormt een beschermende laag die de fijne ondervacht droog houdt. Een paar vierkante centimeter ondervacht bevat tot een miljoen haren. Jammer genoeg heeft dit essentiële kenmerk bijna tot hun uitsterven geleid, doordat commerciële jagers het aantal zeeotters drastisch teruggebracht hebben.

Appendix A.2.1 Vocabulary support given in Chapter 5

Vocabulary support for *The Sun*

(to) **convert**: omzetten

**Dwarf**: dwerg

**Dwarf star**: dwergster, type ster

**Eruption**: uitbarsting

(to) **expand**: uitzetten, verbreden

**Helium**: zonnegas, chemisch element

**Hydrogen**: waterstof, chemisch element

**Mercury**: Mercurius (planeet)

(to) **predict**: voorspellen

**Thermonuclear reaction**: thermonucleaire reactie, kernreactie door hitte

Vocabulary support for *De zon*

**Dwerg-**: klein

**Dwergster**: type ster

**Helium**: zonnegas, chemisch element

**Mercurius**: planeet

**Omzetten**: veranderen in

**Thermonucleaire reactie**: kernreactie door hitte

**Uitbarsting**: eruptie

**Uitzetten**: verbreden

**Voorspellen**: voorspellen

**Waterstof**: hydrogeen, chemisch element
Vocabulary support for Sea Otters

**Abalone**: zeeoor (weekdier)

**Badger**: das (landroofdier)

**Clam**: mossel

**Delicacy**: delicatesse

**Dense**: dicht, compact

(to) **dwell**: vertoeven

(to) **exhaust**: uitputten, opraken

(to) **feast**: verslinden

**Feature**: kenmerk

**Invertebrate**: ongewerveld (dier)

(to) **lack**: ontbreken, niet hebben

**Layer**: laag

**Mammal**: zoogdier

**Nutrient**: voedingsstof

(to) **perish**: verdwijnen, omkomen

**Picky**: kieskeurig

**Sea urchin**: zee-egel (schelpdier)

**Squid**: inktvis

**Voracious**: gulzig

**Weasel**: wezel (landroofdier)
Vocabulary support for Zeeotters

Das: landroofdier

Delicatesse: lekkernij

Dik: compact

Gulzig: vraatzuchtig

Inktvis: weekdier

Kenmerk: eigenschap

Kieskeurig: veeleisend, moeilijk

Mossel: weekdier

Omkomen: sterven

Ongewerveld: zonder benig geraamte

Speklaag: vetlaag

Uitputten: opraken

Verslinden: opvreten

Vertoeven: verblijven

Voedingsstof: nutriënt

Wezel: landroofdier

Zee-egel: stekelhuidig dier

Zeeoor: schelpdier

Zoogdier: dier dat zijn jongen zoogt
Appendix A.2 Tests for the biology texts

Appendix A.2.1 Tests used in Chapter 2

Dutch free recall instruction

*Maak een samenvatting van de tekst die je net las. Vermeld zoveel mogelijk details.*

English free recall instruction

*Write a summary of the text you have just read. Be as detailed as you can be.*

English free recall key for *The Sun* (based on Roediger & Karpicke, 2006)

<table>
<thead>
<tr>
<th>Idea#</th>
<th>Idea</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Sun today is a <strong>yellow dwarf star</strong></td>
</tr>
<tr>
<td>2</td>
<td>It is fueled by <strong>thermonuclear reactions</strong> near its center</td>
</tr>
<tr>
<td>3</td>
<td>The reactions convert hydrogen to helium</td>
</tr>
</tbody>
</table>
| 4     | The Sun has **existed** in its present state for about 4 billion, 600 million years  
(4 billion and some is okay) |
| 5     | And it is **thousands of times larger than the Earth** |
| 6     | **By studying other stars,** |
| 7     | Astronomers can predict **what the rest of the Sun’s life will be like** |
| 8     | About 5 **billion years** from now (billions is okay) |
| 9     | The core of the Sun will **shrink** |
| 10    | And it (the core) will become **hotter** |
| 11    | **The surface temperature will fall** |
| 12    | The higher temperature of the center will **increase the rate of thermonuclear reactions** |
| 13    | The outer regions of the Sun will **expand** |
| 14    | Approximately **55 million kilometers** |
| 15    | Which is about the **distance to Mercury** |
| 16    | The Sun will then be a **red giant star** |
| 17    | **Temperatures on the Earth will become too hot for life to exist** |
| 18    | Once the Sun has used up its **thermonuclear energy** as a red giant |
| 19    | It will begin to **shrink** |
| 20    | After it shrinks to the **size of the Earth** |
| 21    | It will become a **white dwarf star** |
| 22    | The Sun may throw off huge amounts of **gases in violent eruptions** |
23 Called **nova explosions**
24 As it changes from a **red giant to a white dwarf**
25 After **billions of years** as a white dwarf,
26 The Sun will have **used up all its fuel**
27 And it will have **lost its heat**
28 Such a star is called a **black dwarf**
29 After the sun has become a black dwarf, the **Earth will be dark and cold**
30 If any **atmosphere remains there it will have frozen unto the Earth's surface**

English free recall key for *Sea Otters* (based on Roediger & Karpicke, 2006)

<table>
<thead>
<tr>
<th>Idea#</th>
<th>Idea</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sea otters dwell in the <strong>North Pacific</strong></td>
</tr>
<tr>
<td>2</td>
<td>They are the <strong>largest of the mustelids</strong></td>
</tr>
<tr>
<td>3</td>
<td>A group that also includes <strong>freshwater otters, weasels, and badgers</strong></td>
</tr>
<tr>
<td>4</td>
<td>They are from <strong>1 to 1.5 meters long</strong></td>
</tr>
<tr>
<td>5</td>
<td>Most weigh from <strong>25 to 40 kilograms</strong></td>
</tr>
<tr>
<td>6</td>
<td>Large <strong>males</strong> may weigh <strong>45 kilograms or more</strong></td>
</tr>
<tr>
<td>7</td>
<td>Unlike most marine mammals (such as seals or dolphins), sea otters <strong>lack a layer of blubber</strong></td>
</tr>
<tr>
<td>8</td>
<td>Therefore they have to <strong>eat up to 30 percent of their body weight a day</strong></td>
</tr>
<tr>
<td>9</td>
<td>In <strong>clams, crabs, fish, octopus, squids</strong>, and other delicacies</td>
</tr>
<tr>
<td>10</td>
<td>In order to <strong>maintain body heat</strong></td>
</tr>
<tr>
<td>11</td>
<td>Their voracious appetites do <strong>not create food shortages</strong> (no single type of food source is exhausted)</td>
</tr>
<tr>
<td>12</td>
<td>Because they are picky eaters, each animal <strong>preferring only a few food types</strong> (<em>not verbatim</em>)</td>
</tr>
<tr>
<td>13</td>
<td>Sea otters play an important environmental role by <strong>protecting forests of seaweed called kelp</strong></td>
</tr>
<tr>
<td>14</td>
<td><strong>Kelp provide shelter and nutrients</strong> to many species</td>
</tr>
<tr>
<td>15</td>
<td>Certain sea otters <strong>feast on invertebrates (like sea urchins and abalones) that destroy kelp</strong></td>
</tr>
<tr>
<td>16</td>
<td>Sea otters <strong>eat and sleep while floating on their backs</strong></td>
</tr>
<tr>
<td>17</td>
<td>Often <strong>on masses of kelp</strong></td>
</tr>
<tr>
<td>18</td>
<td>They seldom come on shore</td>
</tr>
<tr>
<td>19</td>
<td>Sea otters keep <strong>warm</strong> by means of their luxuriant double-layered fur</td>
</tr>
<tr>
<td>20</td>
<td>Their <strong>fur is the densest among animals</strong></td>
</tr>
<tr>
<td>21</td>
<td>The <strong>soft outer fur forms a protective cover</strong> (protects the underfur)</td>
</tr>
<tr>
<td>22</td>
<td>That <strong>keeps the fine underfur dry</strong></td>
</tr>
<tr>
<td>23</td>
<td><strong>One square inch of underfur contains up to one million hairs</strong></td>
</tr>
<tr>
<td>24</td>
<td>Unfortunately, this essential <strong>feature almost led to their extinction</strong> (endangerment is okay)</td>
</tr>
<tr>
<td>25</td>
<td>Since commercial <strong>hunters</strong> drastically reduced their numbers</td>
</tr>
<tr>
<td></td>
<td>a) <strong>Under government protection.</strong></td>
</tr>
<tr>
<td>26</td>
<td>b) the sea otter population has <strong>recovered</strong> (laws/legal protection)</td>
</tr>
<tr>
<td>27</td>
<td>For example, in <strong>1989</strong></td>
</tr>
<tr>
<td>28</td>
<td>Up to <strong>5,000 sea otters perished</strong></td>
</tr>
<tr>
<td></td>
<td>a) When the <strong>Exxon Valdez</strong></td>
</tr>
<tr>
<td>29</td>
<td>b) <strong>spilled oil</strong></td>
</tr>
<tr>
<td></td>
<td>a) <strong>In Prince William Sound</strong></td>
</tr>
<tr>
<td>30</td>
<td>b) <strong>Alaska</strong></td>
</tr>
</tbody>
</table>
### Tick the correct answer box for every statement, based on the text you have read.

<table>
<thead>
<tr>
<th>Statement</th>
<th>TRUE</th>
<th>FALSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>The conversion of hydrogen into helium happens in all dwarf stars</td>
<td></td>
<td>31</td>
</tr>
<tr>
<td>There are white, yellow and black dwarf stars</td>
<td></td>
<td>32</td>
</tr>
<tr>
<td>The surface of a red giant star is hotter than that of a yellow dwarf star</td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>Temperature on Earth will first increase and then decrease</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>The Sun will exist as a star for more than another 6 billion years</td>
<td></td>
<td>35</td>
</tr>
<tr>
<td>Thermonuclear reactions are temperature-dependent</td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>Dwarf stars are always smaller than planets</td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>In 2 billion years, there will be no life on Earth</td>
<td></td>
<td>38</td>
</tr>
<tr>
<td>White dwarfs are hot stars</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>Temperature on Earth depends on the temperature of the Sun</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>The Sun has been a white dwarf for billions of years</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>When the Sun shrinks, it will become a yellow dwarf star</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>The Earth might get a frozen atmosphere</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>The core of the Sun will become hotter</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>The Sun will shrink to the size of the Earth</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>The Sun uses thermonuclear energy as a red dwarf</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>The core of the Earth will shrink</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>The outer regions of the Sun will expand</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>The Sun will use up all its fuel</td>
<td></td>
<td>26</td>
</tr>
<tr>
<td>The higher temperature in the center of the Sun will increase the rate of thermonuclear reactions</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>The Sun may throw off huge amounts of light in violent eruptions</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>The surface temperature of the Sun will increase</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>The reactions in the sun convert hydrogen to helium</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>The distance between the Sun and the Earth will increase</td>
<td></td>
<td>46</td>
</tr>
<tr>
<td>Astronomers are not sure about the Sun's future</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Hydrogen will be converted to helium on Earth</td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>Once the Sun has used up all thermonuclear energy, she will begin to shrink</td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>A star that has lost her heat is called a cold dwarf</td>
<td></td>
<td>42</td>
</tr>
<tr>
<td>In 5 billion years, the Sun will change</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>The Sun has existed in its present state for about 4 billion, 600 million years</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>The Earth will become too hot for life some day</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>The Sun will change from a red giant to a white dwarf</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>Mercury is a cold planet</td>
<td></td>
<td>44</td>
</tr>
<tr>
<td>The Sun will never lose its heat</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>The Earth will become darker and colder in the long term</td>
<td></td>
<td>29</td>
</tr>
<tr>
<td>Astronomers have studied other planets in order to predict the Sun's life</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Today, the sun is a white dwarf star</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>The Sun is fueled by thermonuclear reactions near its center</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
The degree of expansion of the Sun will be much larger than the distance to Mercury 15
The eruptions of the Sun are called nova stars 23
The Sun will never be a red giant 16
The sun is millions of times larger than the Earth 5
The Sun will eventually become a black dwarf 28
The Sun will be a nova star for a little while 45
The Sun will expand for about 55 million light years 14
Nova explosions only happen in red giants 41

Dutch true false test The Sun (with question number in last column)

<table>
<thead>
<tr>
<th>Duid voor iedere zin het juiste antwoord aan, gebaseerd op de tekst die je net las.</th>
<th>JUIST</th>
<th>FOUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>De omzetting van waterstof naar helium gebeurt in alle dwergsterren.</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Er bestaan witte, gele en zwarte dwergsterren</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Een rode reuzenster is aan de oppervlakte warmer dan een gele dwergster.</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>De temperatuur op aarde zal eerst stijgen en dan dalen</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>De zon zal nog meer dan 6 miljard jaar bestaan als ster</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Thermonucleaire reacties zijn temperatuurafhankelijk</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Dwergsterren zijn steeds kleiner dan planeten</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>Over 2 miljard jaar zal er geen leven meer op aarde zijn.</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Witte dwerger zijn warme sterren</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>De temperatuur op aarde hangt af van de temperatuur van de zon</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>De zon is voor miljarden jaren een witte dwerg geweest</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Wanneer de zon krimpt, wordt ze een gele dwergster</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>De aarde zal mogelijk een vastgevroren atmosfeer krijgen</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>De kern van de zon zal warmer worden</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>De zon zal krimpen tot de grootte van de aarde</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>De zon gebruikt thermonucleaire energie als een rode reus</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>De kern van de aarde zal krimpen</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>De buitenste gebieden van de zon zullen uitzetten</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>De zon zal al haar brandstof opgebruiken</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>De hogere temperatuur in het centrum van de zon zal het aantal thermonucleaire reacties doen stijgen</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>De zon kan grote hoeveelheden licht uitstralen in gewelddadige explosies</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>De oppervlaktetemperatuur van de zon zal stijgen</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>De reacties in de zon zetten waterstof naar helium om</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>De afstand tussen de zon en de aarde zal vergroten</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Astronomen zijn onzeker over de toekomst van de zon</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Waterstof zal op aarde omgezet worden naar helium.</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Eens de zon alle thermonucleaire energie opgebruikt heeft, zal ze beginnen te krimpen</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>
APPENDICES

<table>
<thead>
<tr>
<th>Statement</th>
<th>TRUE</th>
<th>FALSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Een ster die haar warmte verloren heeft, noemen we een koude dwerg</td>
<td></td>
<td>42</td>
</tr>
<tr>
<td>Over 5 miljard jaar zal de zon veranderen</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>De zon bestaat in haar huidige toestand al ongeveer 4 miljard, 600 miljoen jaar</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>De aarde wordt ooit onleefbaar warm</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>De zon zal veranderen van een rode reus naar een witte dwerg</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>Mercurius is een koude planeet</td>
<td></td>
<td>44</td>
</tr>
<tr>
<td>De zon zal nooit haar warmte verliezen</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>De aarde wordt op termijn kouder en donkerder</td>
<td></td>
<td>29</td>
</tr>
<tr>
<td>Astronomen bestuderen andere sterren om de levensloop van de zon te voorspellen</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>De zon is vandaag een witte dwergster</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>De zon wordt gevoed door thermonucleaire reacties bij het centrum</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>De mate van uitzetting van de zon zal groter zijn dan de afstand tot Mercurius</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>De uitbarstingen van de zon noemen we novasteren</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>De zon zal nooit een rode reus zijn</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>De zon is miljoenen keren groter dan de aarde</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>De zon wordt uiteindelijk een zwarte dwerg</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>De zon wordt voor korte tijd een novaster</td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>De zon zal ongeveer 55 miljoen lichtjaren uitzetten</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>Nova-explosies komen enkel voor bij rode reuzen.</td>
<td></td>
<td>41</td>
</tr>
</tbody>
</table>

English true false test Sea Otters (with question number in last column)

Tick the correct answer box for every statement, based on the text you have read.  

<table>
<thead>
<tr>
<th>Statement</th>
<th>TRUE</th>
<th>FALSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>The government has taken measures against the sea otter hunt</td>
<td></td>
<td>31</td>
</tr>
<tr>
<td>Seals and dolphins eat less than sea otters</td>
<td></td>
<td>32</td>
</tr>
<tr>
<td>Sea mammals often cause food shortages</td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>Certain types of animals can survive more easily because of sea otters</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>The underfur of sea otters is precious</td>
<td></td>
<td>35</td>
</tr>
<tr>
<td>Hunters collect the fur and teeth of sea otters</td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>The population of sea otters has been threatened more than once</td>
<td></td>
<td>37</td>
</tr>
<tr>
<td>Sea otters can reach the shore</td>
<td></td>
<td>38</td>
</tr>
<tr>
<td>The Exxon Valdez disaster was the biggest threat for sea otters ever</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>Most seamammals have a layer of blubber</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>Sea otters eat up to 50 percent of their body weight every day</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Barely any animals eat seaweed</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>Prince William acted as the protector of sea otters</td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>The fur of sea otters is the densest amongst animals</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Most sea otters live around Alaska</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>Protection by a group of volunteers helped restore the population of sea otters</td>
<td></td>
<td>26</td>
</tr>
</tbody>
</table>
The Exxon Valdez ditched oil
In 1989, a disaster took place
The voracious appetite of sea otters has never created food shortages
A couple of square centimeters of the underfur consists of about a thousand hairs
The sea otter has a kind of protection for its underfur
Sea otters have a layer of blubber
Sea otters eat everything they can find
In the past, the sea otter was threatened with extinction
Certain sea otters are keen on invertebrates
Millions of sea otters died in an oil spill
The outer fur of sea otters keeps the inner fur dry
Sea otters live in the Pacific Ocean
Freshwater otters and badgers belong to the mustelids
Sea otters eat and sleep while floating on their backs
Sea otters are one to one and a half meters long
The oil spill was covered up by the government
At this time, the population of sea otters is estimated at 6000
Large males can weigh up to 60 kilograms or more
Sea otters never go on shore
The presence of sea otters is an advantage for seaweed
Sea otters sleep a lot in order to maintain body heat
Sea otters keep warm by means of their double-layered fur
Sea otters way between 25 and 40 kilograms
Sea otters can lie on kelp
Sea otters often eat clams, shrimps, and squid
Sea otter hunting has reduced the number of sea otters
The hairs of sea otters are a couple of centimeters long
The oil spill happened in Alaska
Sea otters are the smallest among the mustelids
Sea otters and dolphins are enemies

Dutch true false test Sea Otters (with question number in last column)

Duid voor iedere zin het juiste antwoord aan, gebaseerd op de tekst die je net las.

<table>
<thead>
<tr>
<th>Zin</th>
<th>JUIST</th>
<th>FOUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>De regering heeft maatregelen tegen zeeotterjacht genomen</td>
<td></td>
<td>31</td>
</tr>
<tr>
<td>Zeehonden en dolfijnen eten minder dan zeeotters</td>
<td></td>
<td>32</td>
</tr>
<tr>
<td>Zeezoogdieren veroorzaken vaak voedseltekorten.</td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>Dankzij zeeotters kunnen bepaalde diersoorten gemakkelijker overleven</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>De ondervacht van zeeotters is kostbaar</td>
<td></td>
<td>35</td>
</tr>
<tr>
<td>Jagers verzamelen de vacht en tanden van zeeotters</td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>Beschrijving</td>
<td>Pagina</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------------------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>De zeeotterpopulatie is meer dan eens bedreigd geweest</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>Zeeotters kunnen de kust bereiken</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>De Exxon Valdez-ramp was de grootste bedreiging voor zeeotters ooit</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>De meeste zeezooogdieren hebben een speklaag</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Zeeotters eten elke dag tot 50 procent van hun lichaamsgewicht.</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Bijna geen enkel dier eet zeewier</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Prince William trad op als de beschermers van zeeotters</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>De vacht van zeeotters is de dikste onder de dieren</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>De meeste zeeotters leven rond Alaska</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>Door bescherming van een groep vrijwilligers heeft de zeeotterpopulatie zich hersteld</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>De Exxon Valdez dumpte olie</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>Er vond een ramp plaats in 1989</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>De gulzige eetlust van zeeotters heeft nooit voor voedseltekorten gezorgd</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Een paar vierkante centimeter van de ondervacht bestaat uit zo’n duizend haren</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>De zeeotter heeft een soort van bescherming voor zijn ondervacht</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Zeeotters hebben een speklaag</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Zeeotters eten alles wat ze kunnen vinden</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>De zeeotter werd in het verleden met uitsterven bedreigd</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Bepaalde zeeotters zijn verzot op ongewervelde dieren</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Miljoenen zeeotters stierven in een olieramp</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>De buitenvacht van de zeeotter houdt de ondervacht droog</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Zeeotters wonen in de Stille Oceaan</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Zoetwaterotters en dassen behoren tot de marterachtigen</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Zeeotters eten en slapen terwijl ze op hun rug drijven</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Zeeotters zijn één tot anderhalve meter lang</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>De olieramp werd door de overheid in de doofpot gestopt</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Op dit moment wordt de populatie zeeotters op 6000 geschat</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Grote mannetjesdieren kunnen tot 60 kg of meer wegen</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Zeeotters gaan nooit aan land</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>De aanwezigheid van zeeotters is voordelig voor zeewier</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Zeeotters slapen veel om lichaamswarme te behouden</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Zeeotters blijven warm door hun dubbelgelaagde vacht</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Zeeotters wegen tussen de 25 en 40 kilogram</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Zeeotters kunnen op zeewier liggen</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Zeeotters eten vaak mosselen, garnalen, en inktvis</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>De zeeotterjacht heeft het aantal zeeotters gereduceerd</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>De haren van zeeotters zijn een paar centimeter lang</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>De olieramp vond plaats in Alaska</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Zeeotters zijn de kleinste onder de marterachtigen</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Zeeotters en dolfijnen zijn vijanden</td>
<td>42</td>
<td></td>
</tr>
</tbody>
</table>
Appendix A.2.2 Tests used in Chapter 3 (parallel versions)

Note that in these tests, answer options were:
- True
- False
- I don’t know

**English true false test The Sun: version 1**

<table>
<thead>
<tr>
<th>N°</th>
<th>Question</th>
<th>Repeated</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Today, the sun is a white dwarf star</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>7</td>
<td>Astronomers are not sure about the Sun's future</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>8</td>
<td>In 5 billion years, the Sun will change</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>18</td>
<td>The Sun uses thermonuclear energy as a red giant</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>20</td>
<td>The Sun will shrink to the size of the Earth</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td></td>
<td>The Sun can throw off huge amounts of light in violent eruptions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>The Earth will become too hot for life some day</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>14</td>
<td>The Sun will expand for about 55 million light years</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>26</td>
<td>The Sun will use up all its fuel</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td></td>
<td>The Sun will exist as a star for more than another 6 billion years</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>35</td>
<td>White dwarfs are hot stars</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>24</td>
<td>The Sun will never lose its heat</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>27</td>
<td>The Sun will change from a red giant to a white dwarf</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td></td>
<td>The higher temperature in the center of the Sun will increase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>increase the rate of thermonuclear reactions</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td></td>
<td>The conversion of hydrogen into helium happens in all dwarf stars</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>31</td>
<td>A star that has lost her heat is called a cold dwarf</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>42</td>
<td>Hydrogen will be converted to helium on Earth</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>43</td>
<td>Mercury is a cold planet</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>45</td>
<td>The Sun will be a nova star for a little while</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td></td>
<td>The distance between the Sun and the Earth will increase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>De zon zal krimpen tot de grootte van de aarde</td>
<td>no</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

**Dutch true false test The Sun: version 1**

<table>
<thead>
<tr>
<th>N°</th>
<th>Question</th>
<th>Repeated</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>De zon is vandaag een witte dwergster</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td></td>
<td>Astronomen zijn niet zeker over de toekomst van de zon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Over 5 miljard jaar zal de zon veranderen</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>8</td>
<td>De zon gebruikt thermonucleaire energie als een rode reus</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>18</td>
<td>Over 5 miljard jaar zal de zon veranderen</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>20</td>
<td>De zon zal krimpen tot de grootte van de aarde</td>
<td>no</td>
<td>TRUE</td>
</tr>
</tbody>
</table>
De zon kan grote hoeveelheden licht uitstralen in gewelddadige explosies, maar die zijn niet te vergelijken met gewelddadige explosies op aarde.

De aarde wordt ooit onleefbaar warm, maar dit wordt niet verwacht voor een miljoen jaar of meer.

De zon zal ongeveer 55 miljoen lichtjaren uitzetten, waarmee het heelal de zon mogelijk niet zal bereiken.

De zon zal al haar brandstof opgebruiken, maar dit gebeurt niet voor een miljoen jaar, waarmee de zon nog verder zal kunnen branden.

De zon zal nog meer dan 6 miljard jaar bestaan als ster, maar dit wordt niet verwacht voor een miljoen jaar of meer.

Witte dwergen zijn warme sterren, maar deze zijn niet levensbedreigend voor de zon.

De zon zal nooit haar warmte verliezen, maar dit wordt niet verwacht voor een miljoen jaar of meer.

De zon zal veranderen van een rode reus naar een witte dwerg, maar dit gebeurt niet voor een miljoen jaar of meer.

De hogere temperatuur in het centrum van de zon zal het aantal thermonucleaire reacties doen stijgen, maar dit worden niet verwacht voor een miljoen jaar of meer.

De omzetting van waterstof naar helium gebeurt in alle dwergsterren, maar dit worden niet verwacht voor een miljoen jaar of meer.

Een ster die haar warmte verloren heeft, noemen we een koude dwerg, maar dit wordt niet verwacht voor een miljoen jaar of meer.

Waterstof zal op aarde omgezet worden naar helium, maar dit worden niet verwacht voor een miljoen jaar of meer.

Mercurius is een koude planeet, maar dit wordt niet verwacht voor een miljoen jaar of meer.

De zon wordt voor korte tijd een novaster, maar dit wordt niet verwacht voor een miljoen jaar of meer.

De afstand tussen de zon en de aarde zal vergroten, maar dit wordt niet verwacht voor een miljoen jaar of meer.

De zon kan grote hoeveelheden licht uitstralen in gewelddadige explosies, maar die zijn niet te vergelijken met gewelddadige explosies op aarde.

De aarde wordt ooit onleefbaar warm, maar dit wordt niet verwacht voor een miljoen jaar of meer.

De zon zal ongeveer 55 miljoen lichtjaren uitzetten, waarmee het heelal de zon mogelijk niet zal bereiken.

De zon zal al haar brandstof opgebruiken, maar dit gebeurt niet voor een miljoen jaar, waarmee de zon nog verder kan branden.

De zon zal nog meer dan 6 miljard jaar bestaan als ster, maar dit wordt niet verwacht voor een miljoen jaar of meer.

Witte dwergen zijn warme sterren, maar deze zijn niet levensbedreigend voor de zon.

De zon zal nooit haar warmte verliezen, maar dit wordt niet verwacht voor een miljoen jaar of meer.

De zon zal veranderen van een rode reus naar een witte dwerg, maar dit gebeurt niet voor een miljoen jaar of meer.

De hogere temperatuur in het centrum van de zon zal het aantal thermonucleaire reacties doen stijgen, maar dit worden niet verwacht voor een miljoen jaar of meer.

De omzetting van waterstof naar helium gebeurt in alle dwergsterren, maar dit worden niet verwacht voor een miljoen jaar of meer.

Een ster die haar warmte verloren heeft, noemen we een koude dwerg, maar dit wordt niet verwacht voor een miljoen jaar of meer.

Waterstof zal op aarde omgezet worden naar helium, maar dit worden niet verwacht voor een miljoen jaar of meer.

Mercurius is een koude planeet, maar dit wordt niet verwacht voor een miljoen jaar of meer.

De zon wordt voor korte tijd een novaster, maar dit wordt niet verwacht voor een miljoen jaar of meer.

De afstand tussen de zon en de aarde zal vergroten, maar dit wordt niet verwacht voor een miljoen jaar of meer.
### Appendices

<table>
<thead>
<tr>
<th>N°</th>
<th>Question</th>
<th>Repeated</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>Hydrogen will be converted to helium on Earth</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>44</td>
<td>Mercury is a cold planet</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>45</td>
<td>The Sun will be a nova star for a little while</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>46</td>
<td>The distance between the Sun and the Earth will increase</td>
<td>yes</td>
<td>FALSE</td>
</tr>
</tbody>
</table>

### Dutch true false test The Sun: version 2

<table>
<thead>
<tr>
<th>N°</th>
<th>Question</th>
<th>Repeated</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>De zon wordt gevoed door thermonucleaire reacties bij het centrum</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>5</td>
<td>De zon is miljoenen keren groter dan de aarde</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>6</td>
<td>Astronomen bestuderen andere sterren om de levensloop van de zon te voorspellen</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>25</td>
<td>De zon is voor miljarden jaren een witte dwerg geweest</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>21</td>
<td>Wanneer de zon krimpt, wordt ze een gele dwergster</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>23</td>
<td>De uitbarstingen van de zon noemen we novasterren</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>13</td>
<td>De buitenste gebieden van de zon zullen uitzetten</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>15</td>
<td>De mate van uitzetting van de zon zal groter zijn dan de afstand tot Mercurius</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>16</td>
<td>Wanneer de zon krimpt, wordt ze een rode dwergster zijn</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>33</td>
<td>Een rode reuzenster is aan de oppervlakte warmer dan een gele dwergster.</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>37</td>
<td>Dwarfgalaxies zijn steeds kleiner dan planeten</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>3</td>
<td>De reacties in de zon zetten waterstof naar helium om</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>19</td>
<td>Eens de zon alle thermonucleaire energie opgebruikt heeft, zal ze beginnen krimpen</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>32</td>
<td>Er bestaan witte, gele en zwarte dwergsterren</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>42</td>
<td>Een ster die haar warmte verloren heeft, noemen we een koude dwerg</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>43</td>
<td>Waterstof zal op aarde omgezet worden naar helium.</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>44</td>
<td>Mercurius is een koude planeet</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>45</td>
<td>De zon wordt voor korte tijd een novaster</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>46</td>
<td>De afstand tussen de zon en de aarde zal vergroten</td>
<td>yes</td>
<td>FALSE</td>
</tr>
</tbody>
</table>

### English true false test Sea otters: version 1

<table>
<thead>
<tr>
<th>N°</th>
<th>Question</th>
<th>Repeated</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Sea otters are the smallest among the mustelids</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>13</td>
<td>The presence of sea otters is an advantage for</td>
<td>no</td>
<td>TRUE</td>
</tr>
</tbody>
</table>
seaweed
4 Sea otters are one to one and a half meters long  no  TRUE
6 Large males can weigh up to 60 kilograms or more  no  FALSE
28 Millions of sea otters died in an oil spill  no  FALSE
32 Seals and dolphins eat less than sea otters  no  FALSE

A couple of square centimeters of the underfur
23 consists of about a thousand hairs  no  FALSE
16 Sea otters eat and sleep while floating on their backs  no  TRUE
30 The oil spill happened in Alaska  no  TRUE

Protection by a group of volunteers helped restore
26 the population of sea otters  no  FALSE
33 Sea mammals often cause food shortages  no  FALSE
29 The Exxon Valdez ditched oil  no  FALSE
41 Most sea otters live around Alaska  yes  FALSE
42 Sea otters and dolphins are enemies  yes  FALSE

The hairs of sea otters are a couple of centimeters
43 long  yes  FALSE
44 At this time, the population of sea otters is estimated  yes  FALSE
45 Prince William acted as the protector of sea otters  yes  FALSE
46 The oil spill was covered up by the government  yes  FALSE

Dutch true false test *Sea otters: version 1*

<table>
<thead>
<tr>
<th>N°</th>
<th>Question</th>
<th>Repeated</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Zeeotters zijn de kleinste onder de marterachtigen</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td></td>
<td>De aanwezigheid van zeeotters is een voordeel voor zeewier</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Zeeotters zijn één tot anderhalve meter lang</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td></td>
<td>Grote mannetjesdieren kunnen tot 60 kg of meer wegen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Zeeotters zijn één tot anderhalve meter lang</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>6</td>
<td>Grote mannetjesdieren kunnen tot 60 kg of meer wegen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Miljoenen zeeotters stierven door een olielek</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>32</td>
<td>Zeehonden en dolfijnen eten minder dan zeeotters</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td></td>
<td>Een paar vierkante centimeter van de ondervacht</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>bestaat uit zo'n duizend haren</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>16</td>
<td>Zeeotters eten en slapen terwijl ze op hun rug drijven</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>30</td>
<td>De olieramp vond plaats in Alaska</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td></td>
<td>Door bescherming van een groep vrijwilligers heeft de zeeotterpopulatie zich hersteld</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>De olieramp vond plaats in Alaska</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td></td>
<td>Door bescherming van een groep vrijwilligers heeft de zeeotterpopulatie zich hersteld</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Zeezoogdieren veroorzaken vaak voedseltekorten.</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>29</td>
<td>De Exxon Valdez dumpte olie</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>41</td>
<td>De meeste zeeotters leven rond Alaska</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>42</td>
<td>Zeeotters en dolfijnen zijn vijanden</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>43</td>
<td>De haren van zeeotters zijn een paar centimeter lang</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td></td>
<td>Op dit moment wordt de populatie zeeotters op 6000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>geschat</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>45</td>
<td>Prins William trad op als de beschermers van zeeotters</td>
<td>yes</td>
<td>FALSE</td>
</tr>
<tr>
<td>46</td>
<td>De olieramp werd door de overheid in de doofpot</td>
<td>yes</td>
<td>FALSE</td>
</tr>
</tbody>
</table>
English true false test Sea otters: version 2

<table>
<thead>
<tr>
<th>N°</th>
<th>Question</th>
<th>Repeated</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Sea otters weigh between 25 and 40 kilograms</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>3</td>
<td>Freshwater otters and badgers belong to the mustelids</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>14</td>
<td>Barely any animals eat seaweed</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>11</td>
<td>The voracious appetite of sea otters has never created food shortages</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>7</td>
<td>Sea otters have a layer of blubber</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>9</td>
<td>Sea otters often eat clams, shrimps, and squid</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>15</td>
<td>Certain sea otters are keen on invertebrates</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>17</td>
<td>Sea otters can lie on kelp</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>20</td>
<td>The fur of sea otters is the densest amongst animals</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>1</td>
<td>Zeeotters wegen tussen de 25 en 40 kilogram</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>3</td>
<td>Zoetwaterotters en dassen behoren tot de marterachtigen</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>14</td>
<td>Bijna geen enkel dier eet zeewier</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>11</td>
<td>De gulzige eetlust van zeeotters heeft nooit voor voedseltekorten gezorgd</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>7</td>
<td>Zeeotters hebben een vetlaag</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>9</td>
<td>Zeeotters eten vaak mosselen, garnalen, en inktvis</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>15</td>
<td>Bepaalde zeeotters zijn verzet op ongewervelde dieren</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>17</td>
<td>Zeeotters kunnen op zeewier liggen</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>20</td>
<td>De vacht van zeeotters is de dikste onder de dieren</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>31</td>
<td>De regering heeft maatregelen tegen zeeotterjacht</td>
<td>no</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

Dutch true false test Sea otters: version 2

<table>
<thead>
<tr>
<th>N°</th>
<th>Question</th>
<th>Repeated</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Zeeotters wegen tussen de 25 en 40 kilogram</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>3</td>
<td>Zoetwaterotters en dassen behoren tot de marterachtigen</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>14</td>
<td>Bijna geen enkel dier eet zeewier</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>11</td>
<td>De gulzige eetlust van zeeotters heeft nooit voor voedseltekorten gezorgd</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>7</td>
<td>Zeeotters hebben een vetlaag</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>9</td>
<td>Zeeotters eten vaak mosselen, garnalen, en inktvis</td>
<td>no</td>
<td>FALSE</td>
</tr>
<tr>
<td>15</td>
<td>Bepaalde zeeotters zijn verzet op ongewervelde dieren</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>17</td>
<td>Zeeotters kunnen op zeewier liggen</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>20</td>
<td>De vacht van zeeotters is de dikste onder de dieren</td>
<td>no</td>
<td>TRUE</td>
</tr>
<tr>
<td>31</td>
<td>De regering heeft maatregelen tegen zeeotterjacht</td>
<td>no</td>
<td>TRUE</td>
</tr>
</tbody>
</table>
Appendix A.2.3 Tests used in Chapter 5

English free recall instruction

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.

Dutch free recall instruction

Schrijf op wat je je herinnert uit de tekst die je net las. Je hoeft de tekst niet letterlijk (woord voor woord) neer te schrijven, maar vermeld wel zoveel mogelijk informatie.
Dutch free recall key for The Sun

<table>
<thead>
<tr>
<th>Idee#</th>
<th>Idee</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a) De zon is <strong>een gele dwergster</strong></td>
</tr>
<tr>
<td>2</td>
<td>Ze wordt gevoed door <strong>thermonucleaire reacties</strong> rond het centrum</td>
</tr>
<tr>
<td>3</td>
<td><strong>De reacties zetten waterstof om naar helium</strong></td>
</tr>
</tbody>
</table>
| 4     | De zon *bestaat* in haar huidige toestand *ongeveer 4 miljard en 600 miljoen jaar*  
     | *(4 miljard en … is ook goed)* |
| 5     | en is *duizenden keren groter dan de aarde* |
| 6     | Door andere sterren te bestuderen, kunnen astronomen voorspellen **hoe de rest van de zon haar leven er zal uitzien** |
| 7     | Binnen **5 miljard jaar** *(miljarden is ook goed)* |
| 8     | zal de kern van de zon *krimpen* |
| 9     | en zal ze *(de kern)* *warmer* worden |
| 10    | **De oppervlaktetemperatuur zal dalen** |
| 11    | De hogere temperatuur in het centrum zal het *aantal thermonucleaire reacties doen toenemen* |
| 12    | De buitenste gebieden van de zon zullen *uitsetten* |
| 13    | ongeveer **55 miljoen kilometer** |
| 14    | Dit is ongeveer de *afstand tot Mercurius* |
| 15    | a) De zon zal dan een **rode reuzenster** zijn  
     | b) **De temperaturen op aarde zullen te warm worden**  
     | a) om te kunnen leven  
     | b) na een rode reus al *thermonucleaire energie* opgebruikt heeft |
| 16    | Zal ze beginnen *krimpen* |
| 17    | Wanneer ze verandert van een rode reus naar een witte dwerg |
| 18    | Na **miljarden jaren** als een witte dwerg |
| 19    | Nadat ze gekrompen is tot de *grootte van de aarde* |
| 20    | Die we *nova-explosies* noemen |
| 21    | Wanneer ze verandert van een rode reus naar een witte dwerg |
| 22    | *De zon kan grote hoeveelheden gas* afscheiden in *gewelddadige explosies* |
| 23    | De zon kan grote hoeveelheden *gas* afscheiden in *gewelddadige explosies* |
| 24    | *Dit zijn de* *aardoppervlak* |
| 25    | als een witte dwerg |
| 26    | Nadat de zon een zwarte dwerg geworden is, zal de aarde *donker en koud* zijn |
| 27    | Als er daar een *atmosfeer overblijft, zal die vastgevroren zijn aan het aardoppervlak* |
Dutch free recall key for Sea Otters

<table>
<thead>
<tr>
<th>Idee#</th>
<th>Idee</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Zeeotters vertoeven in de Stille Oceaan</td>
</tr>
<tr>
<td>2</td>
<td>Zij zijn de grootste onder de marterachtigen</td>
</tr>
<tr>
<td>3</td>
<td>een groep die ook uit zoekwaterotters, wezels, en dassen bestaat</td>
</tr>
<tr>
<td>4</td>
<td>Ze zijn 1 tot 1,5 meter lang</td>
</tr>
<tr>
<td>5</td>
<td>de meeste wegen tussen 25 en 40 kilogram</td>
</tr>
<tr>
<td>6</td>
<td>Grote mannetjes kunnen 45 kilogram of meer wegen</td>
</tr>
<tr>
<td>7</td>
<td>In tegenstelling tot de meeste zeezoodieren, zoals zeehonden of dolfijnen, hebben zeeotters geen speklaag</td>
</tr>
<tr>
<td>8</td>
<td>Daarom moeten ze tot 30 procent van hun lichaamsgewicht per dag eten</td>
</tr>
<tr>
<td>9</td>
<td>aan mosselen, krab, vis, octopus, inktvis, en andere delicatessen</td>
</tr>
<tr>
<td>10</td>
<td>Om hun lichaamswarmte te behouden</td>
</tr>
<tr>
<td>11</td>
<td>Hun gulzige eetlust veroorzaakt echter geen voedseltekorten (geen enkele soort van voedselbron uitgeput)</td>
</tr>
<tr>
<td>12</td>
<td>omdat ze kieskeurige eters zijn; elk dier verkiest slechts een aantal voedseltypen</td>
</tr>
<tr>
<td>13</td>
<td>Zeeotters spelen een belangrijke rol in het milieu door de bescherming van wouden van een zeezout dat kelp heet</td>
</tr>
<tr>
<td>14</td>
<td>Kelp voorziet een schuilplaats en voedingsstoffen voor vele diersoorten</td>
</tr>
<tr>
<td>15</td>
<td>Bepaalde zeeotters verslinden ongewervelde dieren, zoals zee-egels en zeeoren, die kelp vernietigen</td>
</tr>
<tr>
<td>16</td>
<td>Zeeotters eten en slapen terwijl ze op hun rug drijven</td>
</tr>
<tr>
<td>17</td>
<td>vaak op kelpmassa’s</td>
</tr>
<tr>
<td>18</td>
<td>Ze komen zelden aan land</td>
</tr>
<tr>
<td>19</td>
<td>Zeeotters houden zich warm door middel van hun luxueuze dubbelgelaagde vacht</td>
</tr>
<tr>
<td>20</td>
<td>Hun vacht is de dikste onder de dieren</td>
</tr>
<tr>
<td>21</td>
<td>De zachte buitenvacht vormt een beschermende laag (beschermt de ondervacht)</td>
</tr>
<tr>
<td>22</td>
<td>die de fijne ondervacht droog houdt</td>
</tr>
<tr>
<td>23</td>
<td>Een paar vierkante centimeter ondervacht bevat tot een miljoen haren</td>
</tr>
<tr>
<td>24</td>
<td>Jammer genoeg heeft dit essentiële kenmerk bijna tot hun uitsterven geleid (”bedreigd” is ook goed)</td>
</tr>
<tr>
<td>25</td>
<td>doordat commerciële jagers het aantal zeeotters drastisch teruggebracht hebben</td>
</tr>
<tr>
<td>26</td>
<td>a) Onder bescherming van de regering (wetten/wettelijke bescherming)</td>
</tr>
<tr>
<td></td>
<td>b) heeft de zeeotterpopulatie zich hersteld</td>
</tr>
<tr>
<td>27</td>
<td>In 1989 bijvoorbeeld</td>
</tr>
<tr>
<td>28</td>
<td>Kwamen 5000 zeeotters om</td>
</tr>
<tr>
<td>29</td>
<td>a) toen de Exxon Valdez</td>
</tr>
<tr>
<td></td>
<td>b) olie lekte</td>
</tr>
<tr>
<td>30</td>
<td>a) in Prince William Sound</td>
</tr>
<tr>
<td></td>
<td>b) in Alaska.</td>
</tr>
</tbody>
</table>
Appendix B. Psychology texts/tests (Chapter 4)

Appendix B.1: Chapter 4 Experiment 1 texts

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 bod 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 presentatiertijd 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 zelfhulp cassettes, 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 zelfhulp bandjes 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.2: Chapter 4 Experiment 1 tests

**Zajonc correction key**

1. What does Zajonc and colleagues’ 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 explanations)

2. Do you think the same results would be found in a group of subjects who have to participate to the experiment for their studies, 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
1. Introduction

Suppose that a student learns a mathematical procedure such as how to find the area of a parallelogram. Later, when the student is given a parallelogram problem like the one she has studied, she is able to compute its area correctly. In short, the student shows that she can perform well on a retention test. However, when this student is asked to find the area of an unusually shaped parallelogram, she looks confused and eventually answers by saying, “We haven’t had this yet.” In short, the student shows that she cannot perform well on a transfer test, that is, on applying what she has learned to a novel situation.

This pattern of good retention and poor transfer is commonly observed among school students (Wertheimer, 1959). On routine problems – that is, problems that are like those they have already learned to solve – they excel; on nonroutine problems – i.e., problems that are not like any that they have solved in the past – they fail. Similar examples can be found in other academic domains, including reading and writing. If a goal of education is to promote transfer as well as retention, then this pattern of performance represents a serious challenge to educators. How can students learn in ways that support solving both routine and nonroutine problems?

How can teachers promote the learning of transferable problem solving skills? More than 50 years ago, Max Wertheimer eloquently posed the questions that motivate this article:

Why is it that some people, when they are faced with problems, get clever ideas, make inventions, and discoveries? What happens, what are the processes that lead to such solutions? What can be done to help people to be creative when they are faced with problems? (Luchins & Luchins, 1970: 1).

Although Wertheimer can be credited with posing an important question, he lacked the research methods and cognitive theories to be able to answer it. The responsibility of Wertheimer’s questioning has been passed to educational psychologists who are concerned with the issue of problem solving transfer (Chipman, Segal & Glaser, 1985; Halpern, 1992; Mayer & Wittrock, in press; Nickerson, Perkins & Smith, 1985; Segal, Chipman & Glaser, 1985). Despite success in understanding how to promote routine problem solving using tried-and-true versions of the drill-and-practice method of instruction, the discipline continues to struggle with how to promote nonroutine problem solving.
What does a successful problem solver know that an unsuccessful problem solver does not know? First, research on problem solving expertise (Chi, Glaser & Farr, 1988; Ericsson & Smith, 1991; Mayer, 1992; Smith, 1991; Sternberg & Frensch, 1991) points to the crucial role of domain-specific knowledge, that is, to the problem solver’s skill. For example, some important cognitive skills for the parallelogram problem include the ability to identify the length and width of the parallelogram, and to perform arithmetic computations such as multiplying length times width to find area. An instructional implication of the skill-based view is that students should learn basic problem-solving skills in isolation.

Unfortunately, mastering each component skill is not enough to promote nonroutine problem solving. Students need to know not only what to do, but also when to do it. Therefore, a second ingredient, suggested by research on intelligence (Sternberg, 1985) and on the development of learning strategies (Pressley, 1990), is the ability to control and monitor cognitive processes. This aspect of problem-solving ability is the problem solver’s metaskill. An instructional implication of the metaskill approach is that students need practice in solving problems in context, that is, as part of working in realistic problem-solving settings.

A focus solely on teaching problem solving skill and metaskill is incomplete, because it ignores the problem solver’s feelings and interest in the problem. A third prerequisite for successful problem solving is suggested by recent research on motivational aspects of cognition (Renninger, Hidi & Krapp, 1992; Weiner, 1986), that is, the problem solver’s will. This approach suggests that problem solving skill and metaskill are best learned within personally meaningful contexts, and that the problem solvers may need guidance in their interpretation of success and failure in problem solving.

The theme of this article is that successful problem solving depends on three components – skill, metaskill, and will – and that each of these components can be influenced by instruction. When the goal of instruction is the promotion of nonroutine problem solving, students need to possess the relevant skill, metaskill, and will. Metacognition – in the form of metaskill – is central in problem solving because it manages and coordinates the other components. In this article, I explore each of these three components for successful problem solving.
2. The role of skill in problem solving

Perhaps the most obvious way to improve problem solving performance is to teach the basic skills. The general procedure is to analyze each problem into the cognitive skills needed for solution and then systematically teach each skill to mastery. Although a focus on teaching basic skills may seem to be the most straightforward way to improve problem solving performance, the results of research clearly demonstrate that knowledge of basic skills is not enough. In this section, I explore three approaches to the teaching of basic skills in problem solving that have developed over the years – instructional objectives, learning hierarchies, and componential analysis – and show how each is insufficient when the goal is to promote problem-solving transfer.

Skills as instructional objectives

Sally wishes to learn how to use a new word processing system, so she takes a course. In the course, she learns how to save and open a document, how to move the cursor, how to insert text, how to delete text, and so on. For each skill, she is given a demonstration and then is asked to solve a problem requiring that skill. She continues on a skill until she can perform it without error; then, she moves on to the next skill. In this way she learns each of the basic skills involved in using the word processing package.

The approach taken in this instruction is to break the subject of word processing into component skills, and then to systematically teach each skill to mastery. In this approach, any large task can be broken down into a collection of “instructional objectives.” Each objective is a single skill, such as being able to move the cursor from the end of a document to some point within the document. Bloom et al. (1956) developed a taxonomy of objectives, and programs of mastery learning were developed to insure that students accomplished each instructional objective (Block & Burns, 1976; Bloom, 1976).

Although mastery programs often succeed in teaching of specific skills, they sometimes fail to support problem-solving transfer. For example, Cariello (reported in Mayer, 1987) taught students to use a computer programming language using a mastery or conventional approach. The mastery group performed better than the conventional group on solving problems like those given during instruction, but the conventional group performed better than the mastery group on solving transfer items. Apparently, narrow focus on mastery of specific objectives can restrict the way that students apply what they have learned to new situations.

Skills as components in a learning hierarchy

Pat is learning how to solve three-column subtraction problems such as, $524 - 251 = \_$. First she practices simple subtraction facts (e.g., $5 - 2 = \_$). Then, she moves on to two column subtraction where no borrowing is needed (e.g., $54 - 21 = \_$). Next, she learns to solve two-column subtraction problems involving borrowing (e.g., $52 - 25 = \_$). In short, she learns to carry out the simpler computational procedures before moving on to the more difficult ones.
This instructional episode is based on Gagne’s (1968; Gagne, Mayor, Garstens & Paradise, 1962) conception of learning hierarchies. A learning hierarchy is a task analysis that yields a hierarchy of subtasks involved in any problem-solving task. Validation of a learning hierarchy occurs when it can be shown that students who pass a higher-level task also are able to pass all prerequisite tasks in the hierarchy (White, 1974). Interestingly, students often are able to pass all prerequisite tasks but still fail to pass the corresponding higher-level task. For example, students who are able to subtract single-digit numbers (such as \( 6 - 1 = 5 \) or \( 15 - 9 = 6 \)) and to regroup two-digit numbers as is required in “borrowing” (such as changing \( 75 \) to 6 tens and 15 ones) may not be able to carry out two-column subtraction (such as \( 75 - 19 = \_ \)). In this situation, students possess all the basic skills but still cannot carry out the task; what may be missing is the ability to organize and control the basic skills within the context of solving the higher-level task. Thus, research on learning hierarchies shows that possessing basic skills is a necessary, but not sufficient prerequisite for successfully solving higher-level problems.

Skills as components in information processing

Dan is taking a course to prepare him for college entrance examinations. As part of the training, he learns how to solve analogy problems, such as:

page:book:: room (a. door, b. window, c. house, d. kitchen)

The teacher describes and provides practice for each step in the process of analogical reasoning. First, Dan learns to encode each term: The A term is page, the B term is book, the C term is room, and there are four possible D terms. Second, Dan learns to infer the relation between the A and B term: in this example, page is a part of book. Third, Dan learns to apply the A–B relation to the C–D terms: room is a part of house. Finally, Dan learns to respond: the answer is (c).

This instructional episode is based on a componential analysis of analogical reasoning (Sternberg, 1985; Sternberg & Gardner, 1983). In componential analysis, a reasoning task is broken down into its constituent cognitive processes. For example, to solve an analogy problem, a problem solver needs to engage in the cognitive processes of encoding, inferring, applying, and responding. Training in componential skills, especially inferring and applying, tends to improve students’ problem solving performance (Robins & Mayer, 1993). However, expertise in executing the component processes is not sufficient for problem-solving transfer. Based on a series of studies, Sternberg (1985) concludes that in addition to possessing cognitive components, problem solvers need to know how to orchestrate and control the cognitive components in any problem-solving task. Sternberg uses to term metacomponents to refer to these required metaskills.
Cognitieve, metacognitieve en motivationele aspecten van probleemoplossing

1. Inleiding

Veronderstel dat een leerling een wiskundige procedure leert zoals het berekenen van de oppervlakte van een parallelloogram. Wanneer aan de leerling later een parallellogramprobleem wordt gegeven zoals ze bestudeerd heeft, is ze in staat om de oppervlakte hiervan correct te berekenen. Kortom, de leerling toont dat zij goed kan presteren op een retentietest. Wanneer aan deze leerling echter gevraagd wordt om de oppervlakte van een ongewoon gevormd parallellogram te berekenen, kijkt ze verward en antwoordt uiteindelijk door het volgende te zeggen, “We hebben dit nog niet gezien”. Kortom, de leerling toont dat ze niet goed kan presteren op een overdracht test, dat is, het toepassen van wat ze heeft geleerd op een nieuwe situatie.

Dit patroon van goede retentie en zwakke overdracht wordt algemeen vastgesteld bij schoolgaande leerlingen (Wertheimer, 1959). Op routine problemen – dat zijn problemen zoals diegene die ze al hebben geleerd op te lossen – blinken ze uit; op niet-routinematige problemen – i.e., problemen die niet lijken op diegene die ze in het verleden hebben opgelost – falen ze. Gelijkaardige voorbeelden vindt men in andere leerdomineinen, inclusief lezen en schrijven. Als het promoten van zowel transfer als retentie een doel is van onderwijs, dan vormt dit prestatiepatroon een serieuze uitdaging voor leerkrachten. Op welke manier kunnen leerlingen leren om zowel routine als niet-routinematige problemen op te lossen?

Hoe kunnen leerkrachten het leren van overdraagbare probleemoplossingsvaardigheden bevorderen? Meer dan 50 jaar geleden, stelde een welbespraakte Max Wertheimer volgende vragen die dit artikel motiveren:

Hoe komt het dat sommige mensen, wanneer ze geconfronteerd worden met problemen, slimme ideeën krijgen, uitvindingen maken en ontdekkingen doen? Wat gebeurt er, welke processen leiden tot zulke oplossingen? Wat kan worden gedaan om mensen te helpen om creatief te zijn als ze worden geconfronteerd met problemen? (Luchins & Luchins, 1970: 1).

Hoewel het stellen van een belangrijke vraag kan worden toegewezen aan Wertheimer, miste hij de onderzoeksmethoden en cognitieve theorieën om deze te kunnen beantwoorden. De verantwoordelijkheid van Wertheimer’s vraag werd doorgegeven aan onderwijspsychologen die zich bezig houden met het onderwerp van probleemoplossingsoverdracht (Chipman, Segal & Glaser, 1985; Halpern, 1992; Mayer & Wittrock, in druk; Nickerson, Perkins & Smith, 1985; Segal, Chipman & Glaser, 1985). Ondanks het succes in het begrijpen van de bevordering van routinematige probleemoplossing door het gebruiken van getest-en-goedgekeurde versies van de ‘dril en herhaal’ instructiemethode, blijft de discipline worstelen met de vraag hoe niet-routinematige probleemoplossing bevorderd kan worden.
Wat weet een succesvolle probleemoplosser dat een onsuccesvolle probleemoplosser niet weet?

Vooreerst, onderzoek over expertise in probleemoplossing (Chi, Glaser & Farr, 1988; Ericsson & Smith, 1991; Mayer, 1992; Smith, 1991; Sternberg & Frensch, 1991) wijst op de cruciale rol van domeinspecifieke kennis, dat zijn de vaardigheden van de probleemoplosser. Enkele voorbeelden van belangrijke cognitieve vaardigheden voor het parallellogramprobleem zijn de mogelijkheid om de lengte en de breedte van het parallellogram vast te stellen, en om cijferberekeningen uit te voeren zoals het vermenigvuldigen van de lengte met de breedte om de oppervlakte te vinden. Deze vaardigheidsbenadering heeft als gevolg voor instructie dat leerlingen probleemoplossingsvaardigheden losstaand moeten leren.

Elke vaardigheidscomponent beheersen is helaas onvoldoende om niet-routinematige probleemoplossing te bevorderen. Leerlingen moeten niet enkel weten wat ze moeten doen, maar ook wanneer ze het moeten doen. Daarom is een tweede ingrediënt, voorgesteld in onderzoek over intelligentie (Sternberg, 1985) en over de ontwikkeling van leerstrategieën (Pressley, 1990), de mogelijkheid om cognitieve processen te controleren en op te volgen. Dit aspect van de probleemoplossingsvaardigheid is de metavaardigheid van de probleemoplosser. Een gevolg van de metavaardigheidsbenadering voor instructie is dat leerlingen probleemoplossing moeten oefenen in de context, namelijk als deel van het werken in realistische probleemoplossings situaties.


Het thema van dit artikel is dat succesvolle probleemoplossing van drie componenten afhanger – vaardigheden, metavaardigheden en wil – en dat elk van deze componenten beïnvloed kan worden door instructie. Wanneer het bevorderen van niet-routinematige probleemoplossing het doel van instructie is, dan moeten de leerlingen over de relevante vaardigheden, metavaardigheden en wil beschikken. Metacognitie – in de vorm van metavaardigheden – staat centraal bij probleemoplossing omdat het de andere componenten leidt en coördineert. In dit artikel verken ik elk van deze drie componenten voor succesvol probleemoplossen.

2. De rol van vaardigheden bij probleemoplossing
Misschien de meest voor de hand liggende manier om probleemoplossingsprestatie te verbeteren is het aanleren van de basisvaardigheden. De algemene procedure is om elk probleem te analyseren volgens de cognitieve vaardigheden die nodig zijn voor de oplossing en dan systematisch elke vaardigheid aan te leren tot het beheersingsniveau. Hoewel de nadruk op het aanleren van basisvaardigheden misschien de meest eenvoudige manier lijkt voor het verbeteren van probleemoplossingsprestaties, tonen onderzoeksresultaten duidelijk aan dat de kennis van de basisvaardigheden niet genoeg is. In dit deel verken ik drie benaderingen voor het aanleren van basisvaardigheden voor probleemoplossing die door de jaren heen werden ontwikkeld – leerdoelen, leerhiërarchieën en componentiële analyse – en toon ik aan hoe elk onvoldoende is voor het bevorderen van probleemoplossingsoverdracht.

**Vaardigheden als leerdoelen**

Sally wil een nieuw tekstverwerkingssysteem leren gebruiken, dus volgt ze een cursus. Tijdens de cursus leert ze hoe een document op te slaan en te openen, hoe de cursor te bewegen, hoe tekst in te voegen, hoe tekst te verwijderen, enz. Voor elke vaardigheid, krijgt ze een demonstratie en wordt ze vervolgens gevraagd om een probleem op te lossen dat deze vaardigheid vereist. Ze blijft een vaardigheid oefenen totdat ze het foutloos kan uitvoeren; daarna gaat ze verder met de volgende vaardigheid. Op deze manier leert ze elke basisvaardigheid die gepaard gaat met het gebruik van het tekstverwerkingspakket.

De benadering bij deze instructie is het opdelen van het onderwerp van tekstverwerking in componentvaardigheden, en om daarna systematisch elke vaardigheid aan te leren tot het beheersingsniveau. Bij deze benadering kan elke grote taak opgedeeld worden in een verzameling van “leerdoelen”. Elk doel is een aparte vaardigheid, zoals in staat zijn om de cursor te bewegen van het einde van een document naar een bepaald punt in het document. Bloom e.a. (1956) ontwikkelden een taxonomie van doelen, en programma’s over beheersingsleren werden ontwikkeld om te verzekeren dat leerlingen elk leerdoel bereikten (Block & Burns, 1976; Bloom, 1976).

Hoewel beheersings programma’s er vaak in slaagden om specifieke vaardigheden aan te leren, falen ze soms in het ondersteunen van probleemoplossingsoverdracht. Bijvoorbeeld Cariello (beschreven in Mayer, 1987) leerde leerlingen een computerprogrammeertaal te gebruiken met behulp van een beheersings- of een conventionele aanpak. De beheersingsgroep presteerde beter dan de conventionele groep bij het oplossen van problemen zoals diegene die gegeven werden tijdens de instructie, maar de conventionele groep presteerde beter dan de beheersingsgroep bij het oplossen van overdrachtsitems. Blijkbaar kan de enge focus op beheersing van specifieke doelstellingen de manier waarop leerlingen toepassen wat ze geleerd hebben op nieuwe situaties beperken.

**Vaardigheden als componenten in een leerhiërarchie**

Pat leert drieëncijferige aftrekkproblemen op te lossen zoals $524 - 251 = __$. Eerst oefent ze eenvoudige aftrekkingen (bijvoorbeeld $5 - 2 = __$). Vervolgens schuift ze op naar tweeëncijferige aftrekking waar lenen niet nodig is (bijvoorbeeld, $54 - 21 = __$). Vervolgens leert ze tweeëncijferige aftrekkproblemen met
lenen op te lossen (bijvoorbeeld 52 – 25 = __). Kortom, ze leert de eenvoudigere berekeningen uit te voeren vooral eer de moeilijkere.

Deze instructiegebeurtenis is gebaseerd op Gagne’s (1968; Gagne, Mayor, Garstens & Paradise, 1962) opvattingen over leerhiërarchieën. Een leerhiërarchie is een taakanalyse die een hiërarchie aan subtaiken oplevert, betrokken in elke probleemoplossingstaak. Validatie van een leerhiërarchie gebeurt wanneer het aangetoond kan worden dat leerlingen die in staat zijn om een taak op een hoger niveau te volbrengen ook in staat zijn om alle voorgaande taken in de hiërarchie te volbrengen (White, 1974). Interessant is dat leerlingen vaak in staat zijn om alle voorgaande taken te volbrengen, maar nog steeds niet de overeenkomstige hogere taak kunnen volbrengen. Bijvoorbeeld leerlingen die in staat zijn om eencijferige getallen af te trekken (zoals 6-1 = 5 of 15-9 = 6) en tweecijferige getallen te hergroeperen zoals wanneer nodig bij “lenen” (zoals het wijzigen van 75 in 6 tientallen en 15 eenheden) maar niet in staat zijn om tweecijferige aftrekkingen (zoals 75-19 = __) uit te voeren. In deze situatie, beschikken de leerlingen al over de basisvaardigheden, maar kunnen ze de taak nog steeds niet uitvoeren; wat misschien ontbreekt is de mogelijkheid om de basisvaardigheden te organiseren en te controleren in de context van het oplossen van de hogere orde taak. Dus, onderzoek naar leerhiërarchieën toont aan dat het bezit van basisvaardigheden een noodzakelijke, maar onvoldoende voorwaarde is om hogere orde problemen op te lossen.

**Vaardigheden als componenten bij informatieverwerking**

Dan volgt een cursus om zich voor te bereiden op de ingangsexamens aan de hogeschool. Als deel van de oefening leert hij analoge problemen op te lossen zoals:

**Pagina:boek::kamer (a. deur, b. raam, c. huis, d. keuken)**

De docent beschrijft en voorziet een oefening bij elke stap in het proces van analoog redeneren. Eerst leert Dan elk woord te coderen:  A woord is pagina, B woord is boek, C woord is kamer, en er zijn vier mogelijke D woorden. Ten tweede leert Dan de relatie tussen het A en B woord af te leiden: in dit voorbeeld, is pagina deel van een boek. Ten derde leert Dan om de A-B relatie toe te passen op de C-D woorden: kamer is deel van het huis. Uiteindelijk leert Dan te antwoorden: het antwoord is (c).

Deze instructiegebeurtenis is gebaseerd op een componentiële analyse van analoog redeneren (Sternberg, 1985; Sternberg & Gardner, 1983). In componentiële analyse wordt een redeneertaak opgesplitst in de samengestelde cognitieve processen. Bijvoorbeeld om een analoog probleem op te lossen, moet een probleemoplosser beginnen met de cognitieve processen van coderen, concluderen, toepassen en reageren. Trainen van componentiële vaardigheden, in het bijzonder concluderen en toepassen, blijkt de probleemoplossingsprestatie van leerlingen te verbeteren (Robins & Mayer, 1993). Expertise in het uitvoeren van de deelvaardigheden is echter niet voldoende voor probleemoplossingsoverdracht. Gebaseerd op een reeks studies concludeert Sternberg (1985) dat naast het bezit van cognitieve componenten, probleemoplossers moeten weten hoe de cognitieve componenten aan te wenden en te controleren in elke probleemoplossingstaak. Sternberg gebruikt de term *metacomponenten* om te verwijzen naar deze verworven metavaardigheden.
Appendix B.4: Chapter 4 Experiment 2 test correction key
(original and translation)

OPEN Vragen

1. Leg uit wat een retentietest is en waarom die soms de voorkeur krijgt.
   Analoge test van geoefende vaardigheden (/1).
   Is goed om na te gaan in hoeverre de student het geleerde kent en kan toepassen in de situaties die onderwezen werden (/1).

2. Welk gevolg had de vraag van Wertheimer voor de onderwijspsychologie?
   Men ging onderzoeken hoe men probleemoplossingsoverdracht/niet-routinematige probleemoplossing kan bevorderen. ("men ging op zoek naar manieren om mensen op een creatieve manier met een probleem te laten omgaan" is niet voldoende, want dat was e vraag van Wertheimer. Iets als "de interesse voor metavaardigheden steeg" is wel oké).

3. Wat zijn volgens de auteur de drie noodzakelijke componenten voor het bevorderen van niet-routinematige probleemoplossing?
   a. (relevante) vaardigheden
   b. Metavaardigheden
   c. Wil (motivatie)

4. Wat is de kernboodschap van de tekst?
   Basisvaardigheden volstaan niet voor het kunnen uitvoeren van hogere orde taken (+ onderwijs voorziet enkel in basisvaardigheden)

5. Hoe effectief is het opdelen van taken in een verzameling van leerdoelen voor probleemoplossing?
   Niet effectief: Beheersingsprogramma's slagen er vaak in om specifieke vaardigheden aan te leren, maar ze falen soms in het ondersteunen van probleemoplossingsoverdracht

MEERKEUZEVRAGEN

1. Voor welke test is het leren van probleemoplossing op basis van een leerhiërarchie het nuttigst?
   a. Een retentietest van een niet-routinematig probleem
   b. Een retentietest van een routinematig probleem.
APPENDICES

c. Een overdracht test van een routinematig probleem.
d. Een overdracht test van een niet-routinematig probleem.

   a. Een retentietest
   b. Een parallelle test
   c. Een overdracht test
   d. Een niet-routinematige test

3. Welke auteur ontwikkelde een taxonomie van leerdoelen om te verzekeren dat leerlingen elk leerdoel bereiken?
   A. Gagne
   B. Mayer
   C. Bloom
   D. Sternberg

4. Welke stelling is correct?
   A. De metavaardigheid van de probleemoplosser bestaat uit de mogelijkheid om cognitieve processen te controleren en op te volgen.
   B. Niet-routinematige probleemoplossing wordt bevorderd door het oefenen van retentietests.
   C. Probleemoplossingsvaardigheden worden best aangeleerd in een neutrale klascontext.
   D. De beheersing van basisvaardigheden volstaat om hogere orde problemen op te lossen.

5. Aan welk alternatief draagt domeinspecifieke kennis het meeste bij?
   a. Basisvaardigheden van probleemoplossing
   b. Metavaardigheden van probleemoplossing
   c. Overdrachtsvaardigheden
   d. Retentievaardigheden

6. … verhoudt zich tot … zoals … zich verhoudt tot …
   a. Basisvaardigheden, retentieprobleem; hiërarchische vaardigheden, overdrachtsprobleem
   b. Retentietest, routinematig probleem; overdracht test, niet-routinematig probleem
   c. Cognitie, routinematige probleemoplossing; wil, niet-routinematige probleemoplossing
   d. Leerdoel, retentieprobleem; metacomponent, overdrachtsprobleem

   a. Een studie die aantoont dat studenten die dit programma volgden het doel van wat ze deden niet begrepen.
   b. Een studie die aantoont dat studenten die dit programma volgden zelfs slecht presteerden op de problemen die ze hadden geoefend.
APPENDICES

c. Een studie die aantoont dat studenten die dit programma volgden de structuur van de taak die ze uitvoerden niet begrepen.
d. Een studie die aantoont dat studenten die dit programma volgden slecht presteerden op het oplossing van problemen die ze niet geoefend hadden.

8. Wat was Wertheimers (1959) kritiek op onderwijs op school?
   a. Scholen leren studenten veelbeoefende problemen oplossen, eerder dan nieuwe problemen.
   b. Scholen besteden teveel aandacht aan overdracht tests en negeren het belangen van retentietests.
   c. Basisvaardigheden vormen een hiërarchie, die in haar geheel aangeleerd moeten worden.
   d. Scholen vergeten het belang van motivatie (de wil) in probleemoplossing.

9. Welke rol ziet de auteur van de tekst (Mayer, 1998) weggelegd voor instructie?
   a. Instructie beïnvloedt alle belangrijke componenten van probleemoplossing.
   b. Instructie is vaak schadelijk voor succesvolle probleemoplossing, omdat het te veel focus op de bestaande problemen.
   c. Instructie zou krachtiger zijn als het gebaseerd was op een gedetailleerde hiërarchische componentenanalyse van de problemen die opgelost moeten worden.
   d. Instructie zou zo specifiek als mogelijk moeten zijn, zodat studenten geen problemen meer hebben om de onderwezen problemen op te lossen.

10. Wat is het meest ironische aan de tekst die je bestudeerd hebt?
    a. Omdat de tekst na een paar pagina’s afgebroken werd, leerde je vooral iets over basisvaardigheden, terwijl de tekst net benadrukt dat basisvaardigheden overschat worden.
    b. Hoewel de auteur waarschuwt voor beheersingsprogramma’s, zijn de voorbeelden die hij van deze programma’s geeft overtuigender dan de voorbeelden van programma’s die hij zelf verkiest.
    c. Hoewel de titel van de tekst ‘Cognitieve, metacognitieve, en motivationele aspecten van probleemoplossing” is, denkt de auteur niet dat instructie helpt bij motivationele aspecten.
    d. De auteur bespreekt uitvoerig Wertheimer (1959), hoewel deze auteur een sterke tegenstander van retentietests en routineproblemen.

11. … verhoudt zich tot … zoals … zich verhoudt tot …
    a. Getest-en-goedgekeurde instructie, routinematige probleemoplossing; vaardigheden, interesse in het probleem
    b. Onderwijspsycholoog, overdracht probleemoplossing; Wertheimer, routine probleemoplossing
    c. Wertheimer, overdracht probleemoplossing; onderwijspsycholoog, routinematige probleemoplossing
    d. ‘Dril en herhaal” instructiemethodes, routine probleemoplossing; metavaardigheid, controle en opvolging van cognitieve processen

12. Waarom is oefening in probleemoplossing in context belangrijk?
a. Beheersing van componentvaardigheden is niet voldoende om probleemoplossing in ongeërfonde situaties te bevorderen.
b. Context verhoogt de motivatie van een persoon, zodat de wil om het probleem op te lossen sterker is.
c. Een focus op niet-contextafhankelijk, geïsoleerd leren induceert oppervlakkig leren, zodat de componenten van de basis probleemoplossingsvaardigheden niet goed geoefend zijn.
d. Context plaatst basis probleemoplossingsvaardigheden in een leerhiërarchie, zodat de componenten van de informatieverwerking meer expliciet worden gemaakt.

13. “De algemene procedure is om elk probleem te analyseren volgens de cognitieve vaardigheden die nodig zijn voor de oplossing en dan systematisch elke vaardigheid aan te leren tot het beheersingsniveau.” Welk alternatief over deze stelling is correct volgens Mayer (1998)?
   a. Deze stelling is geldig voor het aanleren van basisvaardigheden, maar niet voor alle aspecten van probleemoplossing.
   b. Als alle scholen het advies uit deze stelling zouden opvolgen, zou probleemoplossing veel verbeterd worden.
   c. Deze stelling onderschat het aanleren van basisvaardigheden.
   d. Deze stelling is in het bijzonder waar voor niet-routinematige taken, die niet gemeten kunnen worden met retentietests.

14. Maggie is aan het worstelen met de lay-out van haar brief en vraagt haar moeder om hulp. Haar moeder komt helpen en zegt dat zulke dingen vaak vervelend zijn omdat de oplossing soms niet de eerste is waar je aan denkt. Samen zoeken ze een oplossing. Welke component van Maggi’s probleemoplossing is haar moeder NIET aan het helpen.
   a. De leerdoelen
   b. De basisvaardigheden in probleemoplossing
   c. De metavaardigheid
   d. Maggie’s motivatie

English Version

OPEN QUESTIONS

1. Explain what a retention test is and for which reason it is sometimes preferred
   Analogical test of trained skills (/1).
   Works well to check how well the student knows the learn information and how well he/she can apply it in the situations that were taught (/1).
2. What was the consequence of Wertheimer’s question for the field of educational psychology?

Educational psychologists started to investigate how problem-solving transfer/nonroutine problem-solving can be promoted (“they went looking for ways to let people deal with a problem in a creative way” is not enough, because that was Wertheimer’s question. Something like “the attention for metaskills increased” is okay).

3. According to the author, what are the three necessary components for non-routine problem solving?
   a. (relevant) skills
   b. Metaskills
   c. Will (motivation)

4. What is the core message of the text you have just read?

Basic skills are not enough for being able to conduct higher order tasks (+ education only takes care of basic skills).

5. How effective is it for problem solving to break up tasks in instructional objectives?

Not effective: mastery programmes often succeed in teaching specific skills, but they sometimes fail in supporting problem solving transfer.

MULTIPLE-CHOICE QUESTIONS

1. For which test is teaching problem solving on the basis of a learning hierarchy the most helpful?
   b. A retention test of a routine problem.
   c. A transfer test of a routine problem.
   d. A transfer test of a nonroutine problem.

2. Vic is learning to become a car mechanic. All week he has been practicing how to change oil in Volkswagen cars. At the end of the week, he is asked to change oil in a Volvo car while the instructor is watching him. What type of test best describes Vic’s experience at the end of the week according to the text?
   a. A retention test
   b. A parallel test
   c. A transfer test
   d. A nonroutine test

3. Which author developed a taxononie of instructional goals to make sure students reach every instructional goal?
   E. Gagne
   F. Mayer
   G. Bloom
H. Sternberg

4. Which statement is correct?
   E. The metaskill of the problem solver consists of the possibility to control and monitor cognitive processes.
   F. Non-routine problem solving is improved by the training of retention tests.
   G. Problem solving skills are best taught in a neutral classroom context.
   H. The mastery of basic skills is sufficient to solve higher order problems.

5. To which alternative does domain-specific knowledge contribute most?
   a. Basic problem solving skills
   b. Problem solving metaskills
   c. Transfer skills
   d. Retention skills

6. … relates to … as … relates to …
   a. Basic skill, retention problem; hierarchical skill, transfer problem
   b. Retention test, routine problem; transfer test, nonroutine problem
   c. Cognition, routine problem solving; will, nonroutine problem solving
   d. Instructional objective, retention problem; metacomponent, transfer problem

7. Mayer (1998) warns against mastery programs entirely based on instructional objectives. Which evidence does he give?
   a. A study showing that students who followed this program did not understand the goals of what they were doing.
   b. A study showing that students who followed this program even performed badly on the problems they had been practicing.
   c. A study showing that students who followed this program did not understand the structure of the task they were performing.
   d. A study showing that students who followed this program performed badly on solving unpracticed problems.

8. What was Wertheimer’s (1959) critique of school teaching?
   a. Schools teach students to solve well-rehearsed problems rather than new problems.
   b. Schools pay too much attention to transfer tests and neglect the importance of retention tests.
   c. Basic skills form a hierarchy, which must be taught in relation to each other.
   d. Schools overlook the importance of motivation (the will) in problem solving.

9. Which role does the author of the text (Mayer, 1998) see for instruction?
   a. Instruction influences all important components of problem solving.
   b. Instruction is often harmful for successful problem solving, because it focuses too much on the existing problems.
   c. Instruction would be more powerful if it were based on a detailed hierarchical component analysis of the problems to be solved.
d. Instruction should be as specific as possible, so that students have no problems solving the taught problems anymore.

10. What is the most ironic about the text you had to study?
   a. Because the text was cut after a few pages, you mainly learned about basic skills whereas the text warns that these are overrated.
   b. Although the author warns against mastery programs, the examples he gives of these programs are more convincing than the examples he gives of the programs he favors himself.
   c. Although the title of the text is “Cognitive, metacognitive, and motivational aspects of problem solving”, the authors does not think instruction helps for motivational aspects.
   d. The author at length discusses Wertheimer (1959), although this author was a strong proponent of retention tests and routine problems.

11. … relates to … as … relates to …
   a. Tried-and-true practice instruction, routine problem solving; skill, interest in the problem
   b. Educational psychologist, transferable problem solving; Wertheimer, routine problem solving
   c. Wertheimer, transferable problem solving; educational psychologist, routine problem solving
   d. Drill-and-practice method of instruction, routine problem solving; metaskill, control and monitor cognitive processes

12. Why is practice in solving problems in context important?
   a. Mastery of component skills is not enough to promote problem solving in unpracticed situations.
   b. Context increases a person’s motivation, so that the will to solve the problem is stronger.
   c. A focus on non-context dependent, isolated learning induces superficial learning, so that the components of the basic problem-solving skills are not well practiced.
   d. Context puts basic problem skills in a learning hierarchy, so that the components of the information processing are made more explicit.

13. “The general procedure is to analyze each problem into the cognitive skills needed for solution and then systematically teach each component skill to mastery.” Which alternative about this statement is correct according to Mayer (1998)?
   a. This statement is valid for the teaching of basic skills, but does not cover all aspects of problem solving.
   b. If all schools followed the advice given in this statement, problem solving would be much improved.
   c. This statement already underestimates the teaching of basic skills.
   d. This statement is particularly true for nonroutine tasks, which cannot be measured with retention tests.

14. Maggie is struggling with the lay-out of her letter and asks help from her mum. Mum comes and says that these things often are a nuisance because sometimes the solution is not the one
you first think of. Together they search for a solution. Which component of Maggie’s problem solving is mum NOT helping?

a. The instructional objectives
b. The basic problem solving skills.
c. The metaskills
d. Maggie’s motivation
Appendix C: Proficiency tests

Appendix C.1: Dutch multiple-choice vocabulary test
(Chapter 3)


75-item Dutch vocabulary test in a multiple choice format with four answer alternatives. The correct answer is underlined.

1. **Successief**: A. Geslaagd, B. Zegevierend, C. Erfelijk, **D. Achtereenvolgend**
2. **Martelaar**: A. Valsaard, B. Muggenzifter, **C. Lijder**, D. Prutser
3. **Acteur**: A. Beheerder van goederen, **B. Persoon verbonden aan het toneel**, C. Ontwerper van auto’s, D. Functionaris op treinen
4. **Wauwelen**: A. Dromen, B. Schommelen, C. Spelen, **D. Babbelen**
5. **Lenigen**: A. Verzachten, **B. Leegdrinken**, C. Verbuigen, D. Verdedigen
6. **Picaresk**: A. Schilderachtig, B. Met betrekking tot een soldaat, C. Uitbundig, **D. Met betrekking tot een schavuit**
7. **Bretel**: A. Jas, B. Schoen, **C. Broek**, D. Pet
8. **Stagnatie**: A. Stilstand, B. Troonsafstand, C. Wisseling, D. Aanpassing
9. **Schrokop**: A. Domoor, B. Schroothoop, C. Vogelschrik, **D. Gulzigaard**
10. **Knullig**: A. Ontrouw, B. Flauw, **C. Onhandig**, D. Prullerig
11. **Matig**: A. Krachtig blijvend, B. Voordelig blijvend, C. Efficiënt blijvend, **D. Redelijk blijvend**
12. **Droedelen**: A. Doelloos tekenen, B. Betekenisloos mompelen, C. Verknoeien, D. Onbewust besmetten
13. **Divan**: A. Tuinontwerp, **B. Meubelstuk**, C. Auto-onderdeel, D. Operazangeres
14. **Gade**: A. Overtuiging, **B. Echtgenoot**, C. Burgerwacht, D. Klutser
15. **Dignitaris**: A. Munt van een land, **B. Hooggeplaatste ambtenaar**, C. Woestijndier, D. Meerderheidsaandeelhouder
16. **Normatief**: A. Opeenhopend, B. Opbouwend, **C. Dwingend**, D. Mondig
17. **Engerling**: A. Bekrompen man of vrouw, **B. Meikever**, C. Plant, D. Akelige persoon
18. **Riant**: A. Afwijkend, B. Grappig, C. Verzoeningsgezind, **D. Aantrekkelijk**
19. **Onbekwaam**: A. Aanstootgevend, B. Niet passend, **C. Niet geschikt**, D. Niet bezonnen
20. **Paviljoen**: A. Bijgebouw, B. Bijbedoeling, **C. Bijfiguur**, D. Bijgerecht
21. **Facetoog**: A. Trendy café, B. **Insect**, C. Nachtdier, D. Donkerblauw oog
22. **Luit**: A. Bouwmateriaal, B. Dier, C. Keukenapparaat, **D. Muziekinstrument**
23. **Onversaagd**: A. Voortreffelijk, **B. Dapper**, C. Vrijmoedig, D. Oprecht
24. **Weetal**: A. Oneindig groot getal, **B. Betweter**, C. Wijze persoon, D. Klein aantal
25. **Patstelling**: A. Positie van waaruit men kan schieten, **B. Situatie zonder oplossing**, C. Mening die afwijkt, D. Uitspraak van een opschepper
26. **Teint**: A. Specerij, B. Pesterij, **C. Kleur**, D. Gesp
27. **Voorzaat**: A. Gevelornament, B. Ontkiemend zaad, C. Voorouder, D. Schuine afdekking boven een deur
28. **Slaags**: A. In gevecht, B. Roomsgezind, C. Zich door niets onderscheidend, D. Onderdanig
29. **Kakofonie**: A. Geheimschrift, B. Kabaal, C. Vuile praat, D. Signalisatie
30. **Romig**: A. Slaperig, B. Slordig, C. Dik en vloeibaar, D. Met lijm bedekt
31. **Schimpen**: A. Scheuren, B. Schelden, C. Schudden, D. Schuiven
32. **Rups**: A. Hondjes, B. Larve, C. Taartjes, D. Aardig
33. **Opsmuk**: A. Opschudding, B. Versiering, C. Beveiliging, D. Ontplooiing
34. **Laakbaar**: A. Niet te vertrouwen, B. Afkeurenswaard, C. Afschuwwekkend, D. Aan lijden onderhevig
35. **Woelig**: A. Tactvol, B. Turbulent, C. Delicaat, D. Ontroerd
36. **Verguld**: A. Als gunst toegestaan, B. Met smaad bejegend, C. Als voedzaam verkocht, D. Met goud bedekt
37. **Publiekelijk**: A. Bevallig, B. Aansprakelijk, C. Kostbaar, D. Openbaar
38. **Exploitatie**: A. Een niet-democratische staatsvorm, B. Opgeblazenheid, C. Gebruik maken van, D. Loslaten van een orgaan
39. **Masochist**: A. Iemand die graag anderen pijn doet, B. Iemand die geen gezag erkent, C. Iemand die gemakkelijk van mening verandert, D. Iemand die graag vernederd wordt
40. **Ontredderd**: A. In veiligheid, B. Troosteloos, C. Vertederd, D. In gevaar
41. **Relaas**: A. Verslag, B. Troost, C. Steun, D. Familielid
42. **Macaber**: A. Griezelig, B. Kleurrijk, C. Ambitieus, D. Onbetrouwbaar
43. **Grimeren**: A. Beschadigen, B. Beschilderen, C. Beschermen, D. Beschuldigen
44. **Hekelen**: A. Overgieten, B. Spelen, C. Inzouten, D. Bekritiseren
45. **Platvloers**: A. Languit, B. Vlak, C. Grof, D. Effen
46. **Gong**: A. Slaginstrument, B. Sleepinstrument, C. Blaasinstrument, D. Houtinstrument
47. **Perikelen**: A. Rondkijken, B. Slachten, C. Moeilijkheden, D. Aanmoedigen
48. **Exorcisme**: A. Het misbruiken van vertrouwen, B. Het vernielen van cultuurgoederen, C. Het onderdrukken van emoties, D. Het uitdrijven van duivels
49. **Xenofoob**: A. Waterafdrijvend, B. Vreemdelingenhater, C. Iemand met pleinvrees, D. Muziekinstrument
50. **Finesse**: A. Lenigheid, B. Lichaamsconditie, C. Bijzonderheid, D. Levendigheid
51. **Tequila**: A. Schelp, B. Pannenkoekje, C. Monster, D. Alcohol
52. **Verbolgen**: A. Taboe, B. Beduusd, C. Verbluft, D. Boos
53. **Tendencies**: A. Aantrekkelijkheid, B. Neiging, C. Verleiding, D. Bekoring
54. **Priet**: A. Uit overtuiging, B. Tuinhuis, C. Oorspronkelijk, D. Gedeeltelijk
55. **Betichten**: A. Aanvechten, B. Betreuren, C. Bedriegen, D. Aanklagen
56. **Nerf**: A. Marterachtige, B. Bladader, C. Zenuwlijder, D. Sukkel
57. **Guitig**: A. Voordelig, B. Bevorderlijk, C. Plechtig, D. Speels
58. **Stramien**: A. Geheim, B. Moeizaam, C. Patroon, D. Zeer hoog
59. **Wrok**: A. Bouwval, B. Keukengerei, C. Haat, D. Gierigaard
60. **Courant**: A. Vloeiend, B. Gebrekelijk, C. Toegelijkt, D. Te voet
61. **Castagnetten**: A. Fruit, B. Kleren, C. Muziek, D. Groenten
63. **Heling**: A. Aanraken van heilige voorwerpen, B. Aannemen van gestolen goed, C. Aanmanen tot actie, D. Aandrijven van voertuigen

285
65. **Seniel**: A. Breekbaar, **B. Zwakzinnig**, C. Verplaatsbaar, D. Onvast
66. **Veren**: A. Keren, B. Ontdoen, C. Reinigen, **D. Eisen**
67. **Drek**: A. Vocht, B. Lucht, C. Bloed, **D. Mest**
68. **Lijvig**: A. Saai, **B. Dik**, C. Opwindend, D. Lichamelijk
69. **Zeis**: A. Graaien, **B. Maaien**, C. Naaien, D. Zaaien
70. **Rekwisieten**: A. Beperkingen, **B. Benodigdheden**, C. Afbakeningen, D. Versnaperingen
71. **Dorpel**: A. Onverstaanbare spraak, B. Kleine hond, C. Kleine stad, **D. Deur**
72. **Inham**: A. Weiland, B. Nageboorte van een merrie, **C. Baai**, D. Achterbout van een varken
73. **Overstelpen**: A. Overwerken, B. Overhalen, **C. Overladen**, D. Overtreden
74. **Feeks**: A. Schroevendraaier, B. Boor, **C. Tang**, D. Hamer
75. **Dressoir**: A. Werktuig, B. Boom, C. Klimaat, **D. Meubelstuk**

**Appendix C.2: Spelling tests**

**Appendix C.2.1: Dutch spelling test – full version (Chapter 2 & 5 Exp. 2)**

<table>
<thead>
<tr>
<th>van tevoren</th>
<th>onbevreesd</th>
<th>abces</th>
<th>stiekem</th>
<th>diarree</th>
</tr>
</thead>
<tbody>
<tr>
<td>zich neervlijen</td>
<td>calorieën</td>
<td>geëlektrocuteerd</td>
<td>initiatief</td>
<td>eindhalte</td>
</tr>
<tr>
<td>tenisster</td>
<td>niveau</td>
<td>benauwend</td>
<td>slimmeriken</td>
<td>ananas</td>
</tr>
<tr>
<td>authentiek</td>
<td>feliciteren</td>
<td>dreumes</td>
<td>excentriek</td>
<td>bacteriën</td>
</tr>
<tr>
<td>hoofdstad</td>
<td>refrein</td>
<td>guerrillaoorlog</td>
<td>hypnose</td>
<td>interview</td>
</tr>
<tr>
<td>patiënt</td>
<td>exceptioneel</td>
<td>kauwgom</td>
<td>kannibaal</td>
<td>pyjama</td>
</tr>
<tr>
<td>waardeloos</td>
<td>spaghetti</td>
<td>heden ten dage</td>
<td>discipline</td>
<td>chagrijnig</td>
</tr>
<tr>
<td>misleidend</td>
<td>sinaasappelsap</td>
<td>lagedrukgebied</td>
<td>annuleren</td>
<td>geliefkoosd</td>
</tr>
<tr>
<td>reclame</td>
<td>apparaat</td>
<td>geïnteresseerd</td>
<td>alinea</td>
<td>beschouwelijk</td>
</tr>
<tr>
<td>dankzij</td>
<td>cocainemaffia</td>
<td>jaloerie</td>
<td>bulldozer</td>
<td>gefotokopieerd</td>
</tr>
<tr>
<td>molderjefje</td>
<td>licentie</td>
<td>ingrediëntenlijst</td>
<td>broccoli</td>
<td>ternauwenuoord</td>
</tr>
<tr>
<td>ontplooiing</td>
<td>onmiddellijk</td>
<td>solliciteren</td>
<td>sympathie</td>
<td>versjouwen</td>
</tr>
<tr>
<td>februari</td>
<td>hondenhok</td>
<td>hygiëne</td>
<td>blindedarmonsteking</td>
<td></td>
</tr>
<tr>
<td>yoghurt</td>
<td>vergevingsgezind</td>
<td>cadeau</td>
<td>onbemind</td>
<td></td>
</tr>
<tr>
<td>uiteraard</td>
<td>achterwege laten</td>
<td>akkoord</td>
<td>afkickcentrum</td>
<td></td>
</tr>
<tr>
<td>succes</td>
<td>babysitten</td>
<td>museum</td>
<td>neerwaarts</td>
<td></td>
</tr>
<tr>
<td>personage</td>
<td>onverbiddelijk</td>
<td>januari</td>
<td>achterhalen</td>
<td></td>
</tr>
<tr>
<td>agessie</td>
<td>litteken</td>
<td>porselein</td>
<td>leidraad</td>
<td></td>
</tr>
<tr>
<td>allerlei</td>
<td>verbouwereerd</td>
<td>paapeaai</td>
<td>abrupt</td>
<td></td>
</tr>
<tr>
<td>geirriteerd</td>
<td>eczeem</td>
<td>eigenlijk</td>
<td>bruidstaart</td>
<td></td>
</tr>
</tbody>
</table>

**Appendix C.2.2: Dutch spelling test – short version (Chapter 5 Exp. 1)**

<table>
<thead>
<tr>
<th>hoofdstad</th>
<th>allerlei</th>
<th>babysitten</th>
<th>hypnose</th>
</tr>
</thead>
<tbody>
<tr>
<td>waardeloos</td>
<td>onbevreesd</td>
<td>onverbiddelijk</td>
<td>annuleren</td>
</tr>
<tr>
<td>misleidend</td>
<td>niveau</td>
<td>litteken</td>
<td>onbemind</td>
</tr>
<tr>
<td>reclame</td>
<td>feliciteren</td>
<td>benauwend</td>
<td>afkickcentrum</td>
</tr>
<tr>
<td>dankzij</td>
<td>refrein</td>
<td>jaloerie</td>
<td>neerwaarts</td>
</tr>
</tbody>
</table>
Appendix C.2.3 English spelling test – full version (Chapter 2 & 5 Exp.2)

<table>
<thead>
<tr>
<th>acceptable</th>
<th>library</th>
<th>accommodate</th>
<th>ignorance</th>
<th>acquire</th>
</tr>
</thead>
<tbody>
<tr>
<td>accidentally</td>
<td>maintenance</td>
<td>achieve</td>
<td>immediately</td>
<td>acquit</td>
</tr>
<tr>
<td>aggressive</td>
<td>maneuver</td>
<td>acquaintance</td>
<td>incidentally</td>
<td>apparent</td>
</tr>
<tr>
<td>argument</td>
<td>millennium</td>
<td>across</td>
<td>independent</td>
<td>apparently</td>
</tr>
<tr>
<td>believe</td>
<td>neighbor</td>
<td>amateur</td>
<td>intelligence</td>
<td>assassination</td>
</tr>
<tr>
<td>calendar</td>
<td>noticeable</td>
<td>appearance</td>
<td>jewelry</td>
<td>atheist</td>
</tr>
<tr>
<td>cemetery</td>
<td>occurred</td>
<td>beginning</td>
<td>knowledge</td>
<td>basically</td>
</tr>
<tr>
<td>changeable</td>
<td>preferred</td>
<td>business</td>
<td>license</td>
<td>bellwether</td>
</tr>
<tr>
<td>coming</td>
<td>pronunciation</td>
<td>category</td>
<td>miniature</td>
<td>bizarre</td>
</tr>
<tr>
<td>completely</td>
<td>receive</td>
<td>column</td>
<td>occasionally</td>
<td>colleague</td>
</tr>
<tr>
<td>equipment</td>
<td>reference</td>
<td>curiosity</td>
<td>persistent</td>
<td>collectible</td>
</tr>
<tr>
<td>experience</td>
<td>relevant</td>
<td>discipline</td>
<td>piece</td>
<td>committed</td>
</tr>
<tr>
<td>friend</td>
<td>sense</td>
<td>exceed</td>
<td>politician</td>
<td></td>
</tr>
<tr>
<td>glamorous</td>
<td>surprise</td>
<td>existence</td>
<td>principal</td>
<td></td>
</tr>
<tr>
<td>guarantee</td>
<td>tattoo</td>
<td>familiar</td>
<td>religious</td>
<td></td>
</tr>
<tr>
<td>guard</td>
<td>therefore</td>
<td>finally</td>
<td>restaurant</td>
<td></td>
</tr>
<tr>
<td>height</td>
<td>tongue</td>
<td>government</td>
<td>rhythm</td>
<td></td>
</tr>
<tr>
<td>honorary</td>
<td>truly</td>
<td>grateful</td>
<td>tomorrow</td>
<td></td>
</tr>
<tr>
<td>indispensable</td>
<td>weather</td>
<td>happened</td>
<td>tyranny</td>
<td></td>
</tr>
<tr>
<td>interrupt</td>
<td>which</td>
<td>humorous</td>
<td>unfortunately</td>
<td></td>
</tr>
</tbody>
</table>

Appendix C.2.4: English spelling test – short version (Chapter 5 Exp.1)

<table>
<thead>
<tr>
<th>acceptable</th>
<th>neighbor</th>
<th>curiosity</th>
<th>politician</th>
</tr>
</thead>
<tbody>
<tr>
<td>argument</td>
<td>noticeable</td>
<td>exceed</td>
<td>principal</td>
</tr>
<tr>
<td>believe</td>
<td>pronunciation</td>
<td>existence</td>
<td>religious</td>
</tr>
<tr>
<td>changeable</td>
<td>reference</td>
<td>familiar</td>
<td>restaurant</td>
</tr>
<tr>
<td>completely</td>
<td>relevant</td>
<td>finally</td>
<td>tomorrow</td>
</tr>
<tr>
<td>equipment</td>
<td>sense</td>
<td>happened</td>
<td>acquire</td>
</tr>
<tr>
<td>experience</td>
<td>tattoo</td>
<td>intelligence</td>
<td>basically</td>
</tr>
<tr>
<td>friend</td>
<td>which</td>
<td>knowledge</td>
<td>bellwether</td>
</tr>
<tr>
<td>guard</td>
<td>accommodate</td>
<td>license</td>
<td>collectible</td>
</tr>
<tr>
<td>interrupt</td>
<td>beginning</td>
<td>miniature</td>
<td></td>
</tr>
<tr>
<td>library</td>
<td>business</td>
<td>persistent</td>
<td></td>
</tr>
<tr>
<td>maintenance</td>
<td>category</td>
<td>piece</td>
<td></td>
</tr>
</tbody>
</table>
## Appendix D: Questionnaires (motivation & texts)

### Appendix D.1 Motivation and text questionnaire Chapter 2

**Omcirkel voor elke vraag het cijfer dat meest van toepassing is**

**Hoe belangrijk vond je het om goed te scoren op deze test?**

<table>
<thead>
<tr>
<th>Niet belangrijk</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>erg belangrijk</th>
</tr>
</thead>
</table>

**Hoe hoog schat je je eigen resultaat in tegenover dat van je medestudenten?**

<table>
<thead>
<tr>
<th>Heel laag</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>heel hoog</th>
</tr>
</thead>
</table>

**Hoe graag lees je in het Nederlands?**

<table>
<thead>
<tr>
<th>Helemaal niet graag</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>heel graag</th>
</tr>
</thead>
</table>

**Hoe graag lees je in het Engels?**

<table>
<thead>
<tr>
<th>Helemaal niet graag</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>heel graag</th>
</tr>
</thead>
</table>

**In hoeverre kende je de inhoud van de tekst over de zon al vooraf?**

<table>
<thead>
<tr>
<th>Helemaal niet</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>tot in detail</th>
</tr>
</thead>
</table>

**Hoe interessant vond je deze tekst?**

<table>
<thead>
<tr>
<th>Niet interessant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>heel interessant</th>
</tr>
</thead>
</table>

**Hoe moeilijk vond je de inhoud van deze tekst?**

<table>
<thead>
<tr>
<th>Heel gemakkelijk</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>heel moeilijk</th>
</tr>
</thead>
</table>

**Hoe moeilijk vond je de structuur van deze tekst?**

<table>
<thead>
<tr>
<th>Heel gemakkelijk</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>heel moeilijk</th>
</tr>
</thead>
</table>

**In hoeverre kende je de inhoud van de tekst over zeeotters al vooraf?**

<table>
<thead>
<tr>
<th>Helemaal niet</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>tot in detail</th>
</tr>
</thead>
</table>

**Hoe moeilijk vond je de inhoud van deze tekst?**

<table>
<thead>
<tr>
<th>Heel gemakkelijk</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>heel moeilijk</th>
</tr>
</thead>
</table>

**Hoe moeilijk vond je de structuur van deze tekst?**

<table>
<thead>
<tr>
<th>Heel gemakkelijk</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>heel moeilijk</th>
</tr>
</thead>
</table>

**Hoe interessant vond je deze tekst?**

<table>
<thead>
<tr>
<th>Niet interessant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>heel interessant</th>
</tr>
</thead>
</table>
Note that these questionnaires contain codes that can be used to link these questions to variables in the datafiles or codebook, but were not visible to the subjects.

### Leesmotivatie

| [ ] Selecteer de optie die het beste overeenkomt met jouw inschatting. ID: 639928X17047X205525 Vraagcode: importance * |
|---|---|---|---|---|---|---|---|
| Kies het toepasselijke antwoord voor elk onderdeel: |
| Komt niet overeen (1) | Komt weinig overeen (2) | Komt een beetje overeen (3) | Komt matig overeen (4) | Komt sterk overeen (5) | Komt bijna exact overeen (6) | Komt exact overeen (7) |
| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

(039928X17047X205525testimp)

| [ ] Selecteer de optie die het beste overeenkomt met jouw inschatting. ID: 639928X17047X205526 Vraagcode: perf * |
|---|---|---|---|---|---|---|
| Kies het toepasselijke antwoord voor elk onderdeel: |
| Hoe hoog schat je je eigen resultaat in tegenover dat van je medestudenten? (039928X17047X205526perf) |
| Laag (1) | Redelijk laag (2) | Ondergemiddeld (3) | Gemiddeld (4) | Bovengemiddeld (5) | Redelijk hoog (6) | Hoog (7) |
| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
Selecteer de optie die het beste overeenkomt met jouw gevoel.
ID: 639928X17047X205524
Vraagcode: rm *

Kies het toepasselijke antwoord voor elk onderdeel:

<table>
<thead>
<tr>
<th>Komb</th>
<th>Komb</th>
<th>Komb</th>
<th>Komb</th>
<th>Komb</th>
<th>Komb</th>
<th>Komb</th>
</tr>
</thead>
<tbody>
<tr>
<td>niet overeen</td>
<td>weinig overeen</td>
<td>beetje overeen</td>
<td>matig overeen</td>
<td>sterk overeen</td>
<td>bijna exact overeen</td>
<td>exact overeen</td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
</tr>
</tbody>
</table>

Ik vind het fijn om ontspannende teksten (fictie, ondertitels, ...) te lezen in het Nederlands (639928X17047X205524rmned)

Ik vind het fijn om ontspannende teksten (fictie, ondertitels, ...) te lezen in het Engels (639928X17047X205524rmfeng)

Ik vind het fijn om in het kader van mijn opleiding academische teksten te lezen in het Nederlands (639928X17047X205524rmnfen)

Ik vind het fijn om in het kader van mijn opleiding academische teksten te lezen in het Engels (639928X17047X205524rmfeng)

Ik vind het zinvol om in het kader van mijn opleiding academische teksten te lezen in het Nederlands (639928X17047X205524rsmeseeng)

Ik vind het zinvol om in het kader van mijn opleiding academische teksten te lezen in het Engels (639928X17047X205524rsseeng)

Ik vind het interessant om in het kader van mijn opleiding academische teksten te lezen in het Nederlands (639928X17047X205524rmnined)

Ik vind het interessant om in het kader van mijn opleiding academische teksten te lezen in het Engels (639928X17047X205524rminteng)
Selecteer de optie die het beste overeenkomt met jouw gevoel.
ID: 639928X17047X205528
Vraagcode: rmbis *

Kies het toepasselijke antwoord voor elk onderdeel.

<table>
<thead>
<tr>
<th>Onderdeel</th>
<th>Komt niet overeen (1)</th>
<th>Komt weinig overeen (2)</th>
<th>Komt een beetje overeen (3)</th>
<th>Komt matig overeen (4)</th>
<th>Komt sterk overeen (5)</th>
<th>Komt bijna exact overeen (6)</th>
<th>Komt exact overeen (7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ik investeer veel moeite om in het kader van mijn opleiding academische teksten te lezen in het Nederlands (639928X17047X205528rmeffned)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Ik investeer veel moeite om in het kader van mijn opleiding academische teksten te lezen in het Engels (639928X17047X205528rmfoceng)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Ik kan me goed concentreren wanneer ik in het kader van mijn opleiding academische teksten lees in het Nederlands (639928X17047X205528rmfocned)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Ik kan me goed concentreren wanneer ik in het kader van mijn opleiding academische teksten lees in het Engels (639928X17047X205528rmfocned)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Ik begrijp de inhoud wanneer ik in het kader van mijn opleiding academische teksten lees in het Nederlands (639928X17047X205528rmcomned)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Ik begrijp de inhoud wanneer ik in het kader van mijn opleiding academische teksten lees in het Engels (639928X17047X205528rmcomeng)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Ik voel me verplicht om in het kader van mijn opleiding academische teksten te lezen in het Nederlands (639928X17047X205528rmextned)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Ik voel me verplicht om in het kader van mijn opleiding academische teksten te lezen in het Engels (639928X17047X205528rmexteng)</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>
Selecteer de optie die het beste overeenkomt met jouw gevoel.
ID: 639928X17047X205527
Vraagcode: opinion *

Kies het toepasselijke antwoord voor elk onderdeel

<table>
<thead>
<tr>
<th></th>
<th>Komt niet overeen (1)</th>
<th>Komt weinig overeen (2)</th>
<th>Komt een beetje overeen (3)</th>
<th>Komt matig overeen (4)</th>
<th>Komt sterk overeen (5)</th>
<th>Komt bijna exact overeen (6)</th>
<th>Komt exact overeen (7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ik vind dat het gebruik van Engels voor academische teksten in mijn opleiding ontoelaatbaar is (639928X17047X205527op1)</td>
<td>● ●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Ik vind dat het gebruik van Engels voor academische teksten in mijn opleiding verantwoord is (639928X17047X205527op2)</td>
<td>● ●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Ik vind dat het gebruik van Engels voor academische teksten in mijn opleiding een meerwaarde is (639928X17047X205527op3)</td>
<td>● ●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Ik vind dat het gebruik van Engels voor academische teksten in mijn opleiding onvermijdelijk is (639928X17047X205527op4)</td>
<td>● ●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>
Appendix D.3 Text questionnaire in Chapter 3 and 5

These questions were included in the motivation questionnaire in Chapter 2 (see Appendix D.1). Subjects indicate how much prior knowledge they had about a text, how interesting they found the text, how difficult they found the content and the structure of the text.

---

**Percepie van de tekst**

[Duid het cijfer aan dat het het best overeenkomt met jouw ervaring bij het lezen van de tekst.

Question code: bsub] *

Please choose the appropriate response for each item.

- Totaal niet: 1
- Totaal wel: 7

1. Hoe goed kende je de inhoud van deze tekst al vooraf?
2. Hoe moeilijk vond je de inhoud van deze tekst?
3. Hoe moeilijk vond je de structuur van deze tekst?
4. Hoe interessant vond je de inhoud van deze tekst?
5. Hoe moeilijk vond je de structuur van deze tekst?
Appendix D.4 Additional set of questions in motivation questionnaire of Chapter 5 Experiment 2

The questionnaire in Chapter 5 is the same as in Chapter 3 (see Appendix D.2) plus this set of questions.
Data Storage Fact Sheet (Chapter 2)

Studying texts in L2: testtype
Author: Heleen Vander Beken
Date: 15/06/2018

1. Contact details

1a. Main researcher

- name: Heleen Vander Beken
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: heleen.vanderbeken@ugent.be (until September 2018); heleen.vander.beken@gmail.com

1b. Responsible Staff Member (ZAP)

- name: Marc Brsybaert
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: marc.brysbaert@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

* Reference of the publication in which the datasets are reported:
https://doi.org/10.1017/S1366728917000189

* Which datasets in that publication does this sheet apply to?:
All data.

3. Information about the files that have been stored

-----------------------------------------------
3a. Raw data:

* Have the raw data been stored by the main researcher? [x] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
  - [x] researcher PC
  - [x] research group file server (anonymised)
  - [x] other (specify): printed/written tests in the Henri Dunantlaan 2, 150.023

* Who has direct access to the raw data (i.e., without intervention of another person)?
  - [x] main researcher (including personal information)
  - [x] responsible ZAP (anonymised)
  - [x] all members of the research group (anonymised)
  - [ ] all members of UGent
  - [ ] other (specify): ...

3b. Other files:

* Which other files have been stored?
  - [x] file(s) describing the transition from raw data to reported results. Specify: coding guidelines
  - [x] file(s) containing processed data. Specify: overview files of control variables and memory test ratings, analyses
  - [x] file(s) containing analyses. Specify: test statistics that are not reported in detail in the paper
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [x] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: codebook
  - [x] other files. Specify: test materials

* On which platform are these other files stored?
  - [x] individual PC
  - [x] research group file server
  - [x] other: Open Science Framework https://osf.io/2twzd/ (only anonymized processed data and some guidelines or additional statistics)

* Who has direct access to these other files (i.e., without intervention of another person)?
- [x] main researcher
- [x] responsible ZAP
- [x] all members of the research group
- [] all members of UGent
- [x] other (specify): everyone via OSF (only anonymized processed data and some guidelines or additional statistics)

4. Reproduction

* Have the results been reproduced independently?: [ ] YES / [x] NO

* If yes, by whom (add if multiple):
  - name:
  - address:
  - affiliation:
  - e-mail:

Data Storage Fact Sheet (Chapter 3)

Studying texts in L2: long-term recognition memory
Author: Heleen Vander Beken
Date: 15/06/2018

1. Contact details

1a. Main researcher

- name: Heleen Vander Beken
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: heleen.vanderbeken@ugent.be (until September 2018);
  heleen.vander.beken@gmail.com

1b. Responsible Staff Member (ZAP)

vo.2
DATA STORAGE FACT SHEETS

- name: Marc Brysbaert
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: marc.brysbaert@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

* Reference of the publication in which the datasets are reported:

* Which datasets in that publication does this sheet apply to?:
All data.

3. Information about the files that have been stored

3a. Raw data: printed tests and instructions

* Have the raw data been stored by the main researcher? [x] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
- [x] researcher PC
- [x] research group file server
- [x] other (specify): printed/written tests in the Henri Dunantlaan 2, 150.023

* Who has direct access to the raw data (i.e., without intervention of another person)?
- [x] main researcher
- [x] responsible ZAP
- [x] all members of the research group
- [ ] all members of UGent
- [ ] other (specify):

3b. Other files: processed data

__________________________________________________________________
* Which other files have been stored?
  - [x] file(s) describing the transition from raw data to reported results. Specify: correction keys, logbook
  - [x] file(s) containing processed data. Specify: overview files of control variables and memory test ratings
  - [x] file(s) containing analyses. Specify: overview files of analyses and outputs
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [x] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: codebook
  - [x] other files. Specify: texts and test materials

* On which platform are these other files stored?
  - [x] individual PC
  - [x] research group file server
  - [x] other: Open Science Framework https://osf.io/j8hav/ (only anonymized processed data and some guidelines or additional statistics)

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [x] main researcher
  - [x] responsible ZAP
  - [x] all members of the research group
  - [ ] all members of UGent
  - [x] other (specify): everyone via OSF (only anonymized processed data and some guidelines or additional statistics)

4. Reproduction

* Have the results been reproduced independently?: [ ] YES / [x] NO

* If yes, by whom (add if multiple):
  - name:
  - address:
  - affiliation:
  - e-mail:

vo.2
Data Storage Fact Sheet (Chapter 4)

Studying texts in L2: further investigation of factors
Author: Heleen Vander Beken
Date: 15/06/2018

1. Contact details

1a. Main researcher

- name: Heleen Vander Beken
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: heleen.vanderbeken@ugent.be (until September 2018); heleen.vander.beken@gmail.com

1b. Responsible Staff Member (ZAP)

- name: Marc Brsybaert
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: marc.brysbaert@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies

* Reference of the publication in which the datasets are reported:

* Which datasets in that publication does this sheet apply to?:
All data. Note that the data has been collected in one large experimental set-up together with experiments reported in the paper mentioned below. Raw data on print has been stored for the entire set-up as a whole.

3. Information about the files that have been stored

3a. Raw data:

* Have the raw data been stored by the main researcher? [x] YES / [ ] NO

If NO, please justify:

* On which platform are the raw data stored?
  - [x] researcher PC
  - [x] research group file server
  - [x] other (specify): printed/written tests in the Henri Dunantlaan 2, 150.023

* Who has direct access to the raw data (i.e., without intervention of another person)?
  - [x] main researcher
  - [x] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify):

3b. Other files:

* Which other files have been stored?
  - [x] file(s) describing the transition from raw data to reported results. Specify: coding guidelines
  - [x] file(s) containing processed data. Specify: overview files of control variables and memory test ratings
  - [x] file(s) containing analyses. Specify: overview files and detailed test statistics (not reported in the manuscript)
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [x] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: codebook
  - [x] other files. Specify: test materials
* On which platform are these other files stored?
  - [x] individual PC
  - [x] research group file server
  - [x] other: Open Science Framework https://osf.io/c67ya/ (only anonymized processed data, test materials, some guidelines or additional statistics)

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [x] main researcher
  - [x] responsible ZAP
  - [x] all members of the research group
  - [ ] all members of UGent
  - [x] other (specify): everyone via OSF (only anonymized processed data, test materials, some guidelines or additional statistics)

4. Reproduction

* Have the results been reproduced independently?: [ ] YES / [x] NO

* If yes, by whom (add if multiple):
  - name:
  - address:
  - affiliation:
  - e-mail:

Data Storage Fact Sheet (Chapter 5)

Studying texts in L2: L2 recall cost is not just a production effect
Author: Heleen Vander Beken
Date: 25/06/2018

1. Contact details

1a. Main researcher

- name: Heleen Vander Beken
- address: Henri Dunantlaan 2, 9000 Gent
1b. Responsible Staff Member (ZAP)
-----------------------------------------------------------
- name: Marc Brsybaert
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: marc.brysbaert@ugent.be

If a response is not received when using the above contact details, please send an email to
data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
===========================================================
* Reference of the publication in which the datasets are reported:

* Which datasets in that publication does this sheet apply to?:
All experimental data.

3. Information about the files that have been stored
===========================================================
3a. Raw data:
-----------------------------------------------------------
* Have the raw data been stored by the main researcher? [x] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
  - [x] researcher PC
  - [x] research group file server
  - [x] other (specify): printed/written tests in the Henri Dunantlaan 2, 150.023

* Who has direct access to the raw data (i.e., without intervention of another person)?
  - [x] main researcher
  - [x] responsible ZAP
  - [x] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify):
3b. Other files:

* Which other files have been stored?
  - [x] file(s) describing the transition from raw data to reported results. Specify: coding guidelines
  - [x] file(s) containing processed data. Specify: overview files of control variables and memory test ratings
  - [x] file(s) containing analyses. Specify: ...
  - [ ] file(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [x] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: codebook
  - [x] other files. Specify: test materials

* On which platform are these other files stored?
  - [x] individual PC
  - [x] research group file server
  - [x] other: Open Science Framework https://osf.io/p5b3y/ (only anonymized processed data and some guidelines or additional statistics)

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [x] main researcher
  - [x] responsible ZAP
  - [x] all members of the research group
  - [ ] all members of UGent
  - [x] other (specify): everyone via OSF (only anonymized processed data and some guidelines or additional statistics)

4. Reproduction

* Have the results been reproduced independently?: [ ] YES / [x] NO

* If yes, by whom (add if multiple):
  - name:
  - address:
  - affiliation:
  - e-mail:

v0.2